J. Mol. Biol. (1995) **251**, 448–467

JMB

Prediction of Quaternary Structure of Coiled Coils. Application to Mutants of the GCN4 Leucine Zipper

Michal Vieth[1], Andrzej Kolinski[1,2], Charles L. Brooks, III[1] and Jeffrey Skolnick[1]*

[1]*Departments of Molecular Biology and Chemistry, The Scripps Research Institute 10666 N. Torrey Pines Road La Jolla, CA 92037, USA*

[2]*Department of Chemistry University of Warsaw Pasteura 1, 02-093 Warsaw Poland*

**Corresponding author*

Using a simplified protein model, the equilibrium between different oligomeric species of the wild-type GCN4 leucine zipper and seven of its mutants have been predicted. Over the entire experimental concentration range, agreement with experiment is found in five cases, while in two cases agreement is found over a portion of the concentration range. These studies demonstrate a methodology for predicting coiled coil quaternary structure and allow for the dissection of the interactions responsible for the global fold. In agreement with the conclusion of Harbury *et al.*, the results of the simulations indicate that the pattern of hydrophobic and hydrophilic residues alone is insufficient to define a protein's three-dimensional structure. In addition, these simulations indicate that the degree of chain association is determined by the balance between specific side-chain packing preferences and the entropy reduction associated with side-chain burial in higher-order multimers.

Keywords: GCN4 leucine zipper; multimeric equilibrium; quaternary structure prediction; quaternary structure stability; protein folding simulations

Introduction

The biological importance and inherent structural simplicity of coiled coils have made them the object of increasing attention. They are important structural proteins (Phillips *et al*., 1986) and comprise a key motif of DNA (Ferre-D'Amare *et al*., 1993; O'Shea *et al*., 1991) and RNA (Banner *et al*., 1987; Cohen & Parry, 1986) binding proteins. Their native structure consists of two or more helices wrapped around each other with a left-handed, helical supertwist (Crick, 1953). Coiled coils exhibit a characteristic seven residue repeat $(abcdefg)_n$ (Hodges *et al*., 1981; McLachlan & Stewart, 1975). Positions *a* and *d* are mostly occupied by hydrophobic residues and form the interface between helices. Positions *b*, *c*, *e*, *f* and *g* are hydrophilic, with *g* and *e* being charged (Cohen & Parry, 1990; Hodges *et al*., 1981). Residues occupying the *g* and *e* positions are believed to play a role in defining the mutual orientation of the helices. Furthermore, since coiled coils are the simplest examples of quaternary structure, they represent a very useful model system for exploring the factors responsible for the stability and specificity of oligomeric structures. In this context, Harbury *et al*. (1993) simultaneously substituted the four *a* residues of the GCN4 leucine zipper (Val9, Asn16, Val23 and Val130) and the four *d* residues (Leu5, Leu12, Leu19 and Leu26) by Leu, Ile and Val. All peptides were found to be more than 90% helical, and their oligomerization states were determined by equilibrium ultracentrifugation and gel filtration. The modified peptides were named according to the identity of the residues in the *a* and *d* positions (e.g. LI stands for the mutant with Leu (Ile) in all four of the *a* (*d*) positions). The IL mutant and the wild-type populate dimeric species; II, LL, LV are trimeric, and LI is tetrameric. The VL mutant populates both dimeric and trimeric species, and the VI mutant populates multiple species.

The goal of this paper is to extend our previous predictions of the folding pathway and structure of the wild-type GCN4 leucine zipper (Vieth *et al*., 1994a) to calculate the equilibrium constant between different oligomeric species. The most straightforward method would be to simulate the oligomerization process directly, following the basic ideas used for the prediction of structure and folding pathways of the GCN4 leucine zipper dimer (Vieth *et al*., 1994a). However, the computer time required for the simulation of a number of chains leading to the formation of a oligomerization state greater than dimers is well beyond our computational resources.

Abbreviations used: PDB, Protein Data Bank; r.m.s.d., root-mean-square deviation.

0022–2836/95/330448–20 $12.00/0

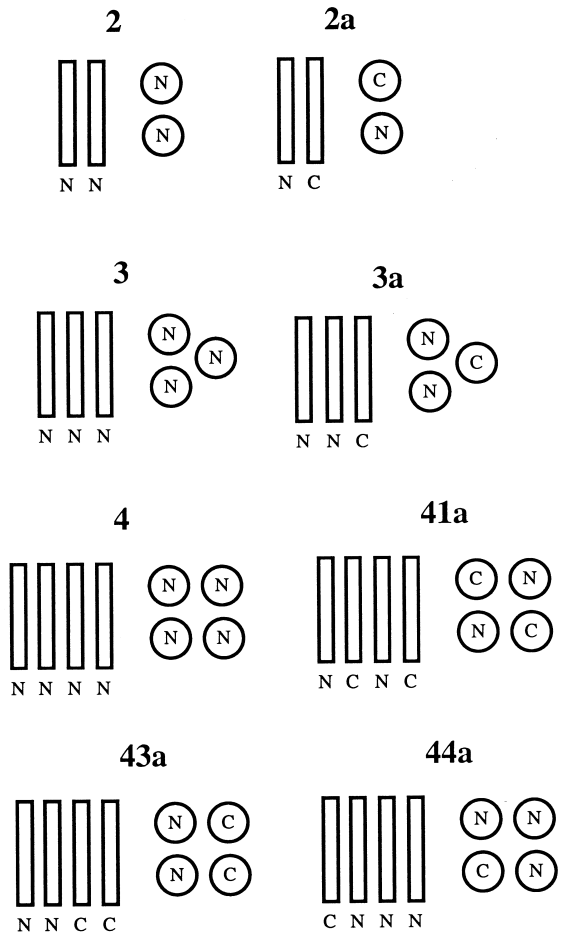

Figure 1. Schematic drawing of the interhelical orientations studied: 2 represents parallel dimers; 2a antiparallel dimers; 3 parallel trimers; 3a antiparallel trimers; 4 parallel tetramers. 41a, 43a, and 44a represent possible antiparallel tetramers studied in this work.

In addition, the folding process would need to be repeated hundreds of times to be statistically significant. Thus, we have opted to develop a methodology that assumes a spectrum of parallel and antiparallel oligomers and attempt to estimate the equilibrium constants within the set of assumed species (schematically shown in Figure 1). The methodology proposed for accomplishing this is based on a new application of the classical Mayer and Mayer statistical mechanical approach (Mayer & Mayer, 1963). The basic idea presented here is to use a computer simulation to obtain the variables necessary for the statistical mechanical treatment. Most importantly, the method presented below allows us to identify, in the context of the model, the dominant interactions responsible for the quaternary structure of coiled coils. To achieve this identification, the method assumes that the energy landscape is such that there are distinct minima corresponding to different, bound oligomerization states (dimers, trimers or tetramers, which may be either parallel or antiparallel) and that the barriers between them are large enough to be considered effectively infinite.

Method

An overview of the entire simulation methodology is presented in Figure 2. The lattice model we use for the estimation of the equilibrium constants is based on an α-carbon representation of the protein backbone and a multiple rotamer, single ball representation of the side-chains (Kolinski & Skolnick, 1994a,b). The entire parameter set for the force field, together with the scaling factors for the different energy terms and their description, is available by anonymous ftp (Vieth *et al*., 1994b).

Lattice model of proteins

Geometric representation and move set

The protein model used for these studies is based on an α-carbon representation of the protein backbone and a single sphere, multiple rotamer representation of side-chains. A schematic representation of the model protein chain is shown in Figure 3. α-Carbon atoms are placed on an underlying cubic lattice with a grid spacing of 1.22 Å and are connected by a set of vectors of the type 1.22* $\{(2,2,0), (2,2,1), (3,0,0), (3,1,0), (3,1,1)\}$ (Kolinski & Skolnick, 1994b). Thus, there are a total of 90 possible ways of choosing a vector connecting two consecutive α-carbon atoms. There are some restrictions for two or three consecutive vectors that prevent non-protein-like conformations of the backbone. Two consecutive C^{α}-C^{α} vectors must have their valence bond angle in the range 72.5° to 154°. Three consecutive C^{α}-C^{α} vectors are allowed if the length of their sum is larger than 4.05 Å. These restrictions lower the total number of three consecutive vector occurrences (90^3) by roughly 60%. The inherent geometric resolution of this lattice is very high; crystal structures from the Brookhaven Protein Data Bank (PDB) can be projected onto the lattice with an average root-mean-square deviation, RMSD of 0.6–0.7 Å (Godzik *et al*., 1993a).

The Monte Carlo move set, depicted schematically in Figure 4, consists of two bond moves, three bond rearrangements, small shifts of larger chain pieces, chain ends modifications and rotamer equilibration. Since the object of the present study is the examination of fluctuations about assembled structures, rigid body shifts of the individual chains (used in the folding of the GCN4 leucine zipper to speed up the assembly process) were not used. One Monte Carlo ''time'' cycle for this system is defined as $(N-2)*M$ two bond moves, $2*M$ two bond moves, M shifts of the chain pieces, and $M*(N-3)$ three bond moves, where M is the number of chains, which varies from 2 to 4 and N is the number of α-carbon atoms. A typical simulation run consisted of 180,000*M cycles.

Statistical potential extracted from high resolution structures from PDB

With two minor exceptions, the interaction scheme used for the lattice model of coiled coils is identical to the one described in a recent paper (Vieth *et al*., 1994a) for the *de novo* folding of GCN4 dimers. A 6 *kT* penalty for too many contacts for a given residue, designed to prevent aggregation in unstructured clusters, was not used here. On examining trimers and tetramers, we found that the number of contacts exceeded the thresholds employed in the folding experiments. However, even when this term is deleted, the folding of the GCN4 leucine zipper from a pair of random chains still occurs, but the folding process is somewhat slowed down. We also abandoned the harmonic

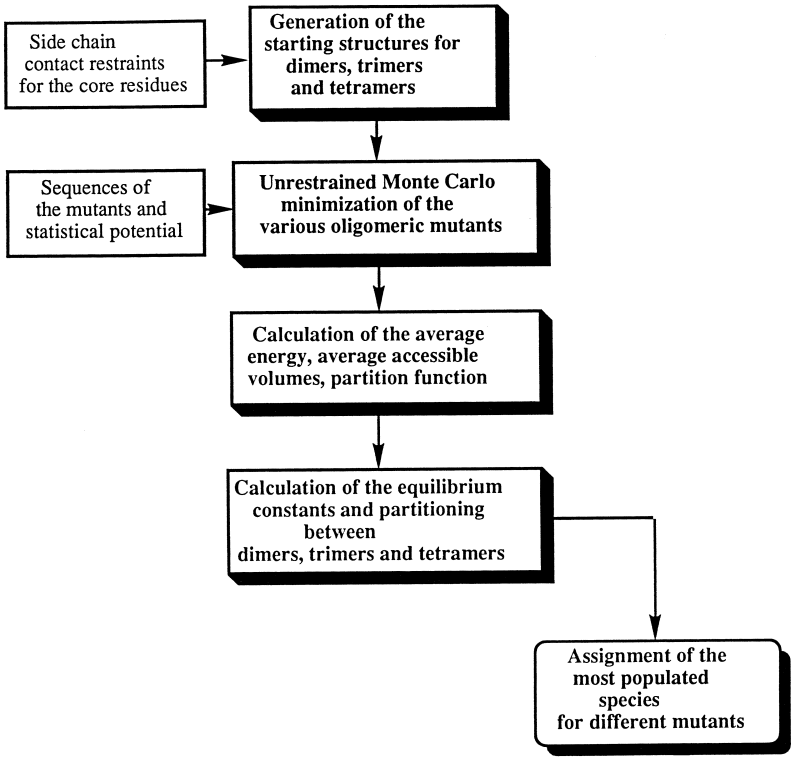

Figure 2. For each mutant, every oligomer is generated and subjected to unrestrained, isothermal Monte Carlo simulations under conditions (energy function, temperature) identical to those for which the wild-type GCN4 leucine zipper was refined (Vieth *et al*., 1994a). Then, the partition functions for each mutant in every oligomeric state are calculated and the most populated species are assigned for the relevant chain concentration.

restraint between centers of masses of the side-chains used to regulate the concentration of chains in the folding experiments.

The interaction scheme is derived in the same spirit as that of Kolinski & Skolnick (1994b), but with some differences and improvements in the energy terms. Apart from a larger database used for the derivation of all of the statistical terms in the potential, the major difference arises from the use of a different burial term (coiled coils do not have spherical geometry; thus, the centrosymmetric burial potential used for single domain globular proteins cannot be used) and a slightly different description of the short range interactions. The local effective Ramachandran potential (that depends on the chirality and the end-to-end distance of three consecutive vectors) is now residue-dependent. In addition, the term describing local side-chain orientations depends on the angular correlation of neighboring C^{α}-C^{β} groups (rather than on the C^{α}-side-chain center of mass orientations used in the original treatment of Kolinski & Skolnick (1994b) for globular proteins). Both changes act to provide sharper interfaces between secondary structural elements. Nevertheless, both factorizations of the local conformational propensities yield very similar results when applied to the folding of Rop monomer, protein A and crambin (Kolinski & Skolnick, 1994a). Furthermore, the backbone-dependent rotamer library used for all coiled coil studies is an improved version of the rotamer library reflecting better statistics in the larger database. Based on the folding of the Hodges' sequences (Hodges *et al*., 1981) and a test of the dynamic stability of assembled dimers, slightly different scaling factors (Kolinski & Skolnick, 1994b) for the different energy terms were chosen to keep the helix content of the non-interacting chains below 50%, as well as to maintain a proper balance of the short-range and long-range interactions. The scaling factors for all of the coiled coils systems studied below are the same as in the previous work on GCN4 leucine zipper folding (Vieth *et al*., 1994a).

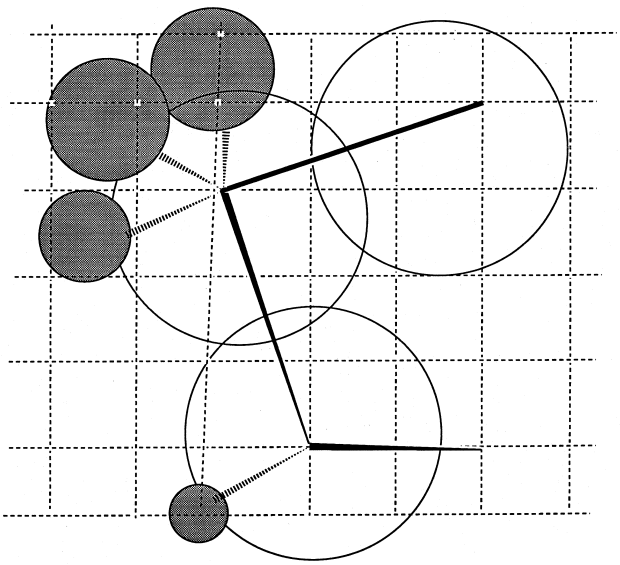

Figure 3. Schematic representation of the model protein chain. α-Carbon atoms are positioned on the underlying cubic lattice with mesh size 1.22 Å and connected by a set of vectors of type $\{(3,1,0)\ (3,1,1),\ (3,0,0),\ (2,2.1),\ (2,2,0)\}$. Each α-carbon has its own repulsive region (white circles). Side-chains are represented by single sphere, multiple rotamers (shaded circles).

The entire interaction scheme can be subdivided into short and long-range interactions. By short-range, we mean local conformational preferences of neighboring residues along the sequence. Similarly, by the term long-range interactions, we mean all interactions that are at least four residues apart. The entire potential (with the exception of the hydrogen bond term) is based on a statistical analysis

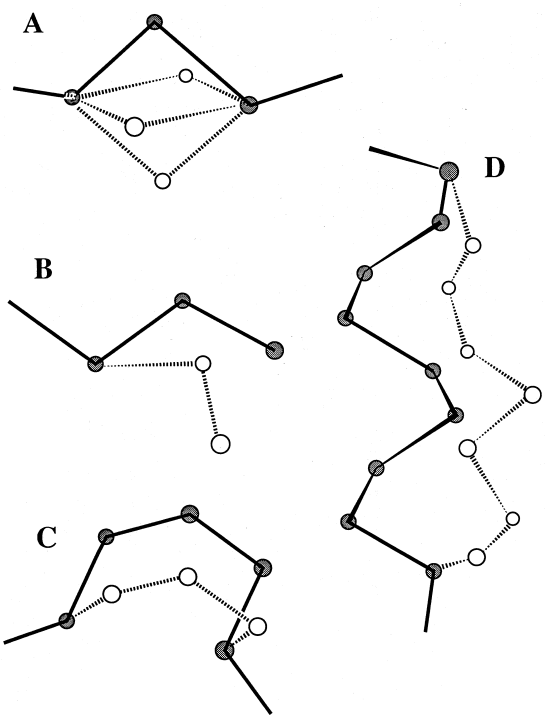

Figure 4. Monte Carlo move set used in this study. Continuous lines and shaded circles represent premodified bonds and α-carbon positions, respectively. Shaded lines and open circles represent modified bonds and α-carbon positions. A, two bond moves; B, chain end modifications; C, three bond moves; D, shifts of the chain pieces.

of a set of high resolution crystal structures from the PDB database. The list of the structures used in the derivation of the potential is provided in Table 1. In what follows, the individual terms are presented in detail. All parameters are available *via* anonymous ftp (Vieth *et al*., 1994b).

Short range interactions: hydrogen bonds

The model hydrogen bond potential is residue-independent and is defined based on the main-chain geometry. Each α-carbon can participate in at most two hydrogen bonds (proline is an exception and can participate in only one hydrogen bond), and there is no directionality (donor-acceptor) in the scheme. Two residues i, j are considered to be hydrogen-bonded if they satisfy the following distance and orientational criteria:

$$|i-j| \geqslant 3$$
$$4.6\ \text{Å} \leqslant |\mathbf{r}_{ij}| \leqslant 7.3\ \text{Å}$$
$$|(\mathbf{b}_{i-1} - \mathbf{b}_i)\cdot\mathbf{r}_{ij}| \leqslant 13.4\ \text{Å}^2$$
$$|(\mathbf{b}_{j-1} - \mathbf{b}_j)\cdot\mathbf{r}_{ij}| \leqslant 13.4\ \text{Å}^2 \qquad (1a)$$

where $\mathbf{r}_{ij}$ is the vector connecting the α-carbons of residues i and j, and the $\mathbf{b}_k$ ($k = i-1, i, j-1, j$) are the corresponding bond vectors which are shown in Figure 5. This model hydrogen bond, which is very similar in spirit to the Levitt-Greer (1977) method of secondary structure assignment, reproduces about 90% of the main-chain hydrogen bonds as assigned by the Kabsch-Sander (1983) program for real proteins. An explicit cooperativity is also introduced into the hydrogen bond scheme. When two neighboring pairs of residues are hydrogen bonded, the system gets an additional favorable energy (cooperativity). The hydrogen bond contribution to the potential can be expressed as follows:

$$E_{\text{HB}} = \sum_{i=1}^{N-3} \sum_{j=1,\, i+3}^{N} (E^{\text{H}}\delta_{ij} + E^{\text{HH}}\delta_{ij}\,\delta_{i+1,\,j+1}) \qquad (1b)$$

where $E^{\text{H}} = -0.6\,kT$ is the hydrogen bond energy, $E^{\text{HH}} = -0.75\,kT$ is the cooperativity energy. $\delta_{ij} = 1(0)$ if residues i and j are (not) hydrogen bonded.

Effective Ramachandran potential

This sequence-specific part of the potential is introduced to correct the lattice distribution of C^{α_i}-$C^{\alpha_{i+3}}$ distances so that they reflect a protein-like distribution of local conformations. Consider a set of three consecutive bond vectors $\mathbf{b}_{i-1}$, $\mathbf{b}_i$ and $\mathbf{b}_{i+1}$; see Figure 6. Then let

$$R_{14}^{i} = |\mathbf{b}_{i-1} + \mathbf{b}_i + \mathbf{b}_{i+1}|^2 X \qquad (1c)$$

where $X = \text{sign}\{(\mathbf{b}_{i-1} \times \mathbf{b}_i)\cdot\mathbf{b}_{i+1}\}$. The potential we employ depends on both R_{14}^{i}, which is amino acid pair-specific as well as a generic propagator term that depends on the

Table 1. PDB codes of the 235 proteins used for the derivation of the statistical potentials

1imm	1ipd	1kst	1l01	1lap	1ldm	1lfg	1lh1	1lhm	1lig	1mbc	1mpp
1mrm	1nn2	1ntp	1ofv	1omd	1paz	1pgd	1pgx	1phh	1pii	1pk4	1ppa
2mhr	1ubq	1ctf	1aai	1c5a	2ovo	1hcc	1gd1	2aza	2fbj	1ppd	1psg
1q21	1r69	1rat	1rbp	1rcb	1rn4	1rnh	1s01	1sgc	1sgt	1snc	1tfd
1tgi	1tie	1ton	1trb	1xis	2aaa	2act	2apr	2bb2	2bus	2c2c	2cdv
2cna	2ctx	2fcr	2fgf	2fx2	2fxb	2gbp	2gcr	2gn5	2lhb	2liv	2nn9
2npx	2por	2prk	2reb	2rhe	2sn3	3adk	3b5c	3bcl	3blm	3cln	3dfr
3enl	3est	3grs	3icb	3icd	3pfk	3wrp	41bi	4fd1	4rxn	4tgf	4tms
4tnc	5acn	5dfr	6rxn	6taa	8adh	1aap	1abm	1acb	1ake	1bbh	1bbk
1bbp	1bbq	1bov	1c2r	1cdt	1cgi	1cob	1col	1cpc	1er8	1fbh	1fcb
1fdl	1fxa	1gct	1gmf	1gp1	1gsr	1gst	1hmd	1hne	1hrh	1il8	1ith
1ldn	1lld	1lmb	1lpr	1mbl	1mca	1msb	1mvp	1nbt	1nsb	1ova	1ovo
1pbx	1pp2	1prc	1rnb	1rve	1sar	1sdh	1sdy	1tab	1tec	1tlp	1tpk
1vaa	1wgc	256b	2ccy	2ci2	2fb4	2hip	2hla	2ltn	2scp	2tim	2tpr
2trx	2utg	2ypi	3fis	3gap	3rp2	3sdp	3sic	4cpa	4hhb	4mdh	4phv
4sgb	4tsl	5rub	8atc	8cat	1aak	1aaj	1aba	1alb	1alc	1ald	1ama
1apb	1aps	1apt	1atx	1avr	1bbc	1bia	1bp2	1bti	1ca2	1cbx	1ccp
1ccr	1cd4	1cdp	1cla	1cms	1cox	1cp4	1csc	1drl	1dri	1dtx	1eca
1ego	1end	1epg	1ezm	1f3g	1fha	1fkf	1fxd	1gal	1gky	1gly	1gox
1gpr	1gps	1hbg	1hid	1hoe	1ifb	1rop					

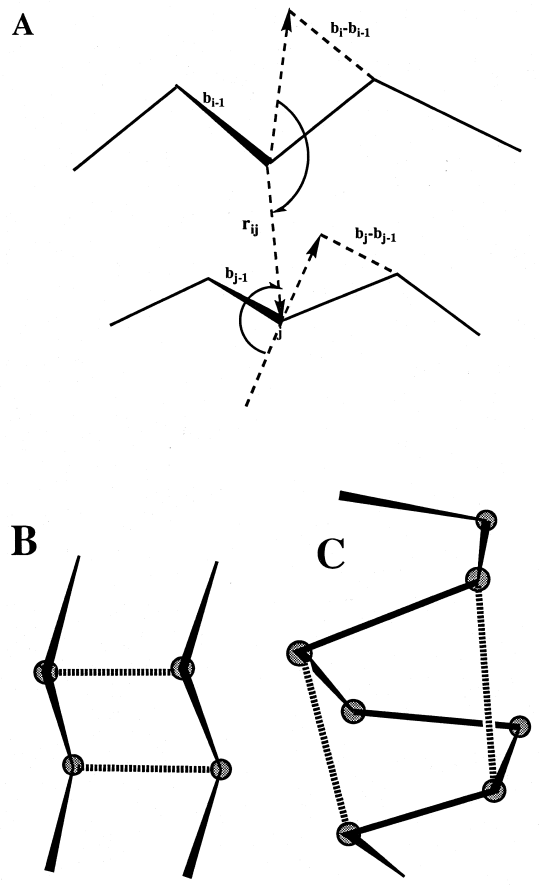

Figure 5. Schematic representation of the hydrogen bond scheme. Geometric criteria used for assessment of whether residues i and j are hydrogen bonded. $\mathbf{b}_{i-1}$ ($\mathbf{b}_{j-1}$) represents bond vectors connecting residues $i-1$ and i ($j-1$ and j). r_{ij} is the distance between α-carbon atoms of ith and jth residues (see equation (1a)). Broken lines represent hydrogen bonds between extended pieces of the chain. Shaded lines represent helical hydrogen bonds.

depends on the conformation states of R_{14}^{i-1} and R_{14}^{i}, but because of poor statistics, this term is amino acid-independent.

$$E_{R_{14}} = \sum_{i=2}^{N-2} \left(\begin{array}{l} 3E_{14}(\mathrm{R}_{14}^{i}, A_i, A_{i+1}) \\ + E_{14|14}(R_{14}^{i-1} | R_{14}^{i}) \end{array} \right) \qquad (1d)$$

A_i is the type of amino acid located at residue i. For each pair of amino acids, E_{14} is divided into 12 discrete bonds, whose relationship to more standard secondary assignments is summarized in Table 2A. Similarly, the residue-independent coupling energy $E_{14|14}$, whose values are presented in Table 2B, is a 12×12 asymmetric matrix that describes the average energy of R_{14}^{i} given that it has been preceded by R_{14}^{i-1}.

Side-chain rotamer energy

A side-chain rotamer library has been constructed which is similar in spirit to that presented by Kolinski & Skolnick (1994b). With the obvious exception of glycine, each side-chain is represented by a single point located at the side-chain center of mass. The set of allowable conformations for the side-chains associated with the ith C^{α} depends on the virtual bond angle between backbone bonds $\mathbf{b}_{i-1}$ and $\mathbf{b}_i$, θ_i. Different residues have a different number of rotamers, whose number ranges from one for alanine to 21 for some backbone conformations of arginine. The maximum number of rotamers for different residues are shown in Table 3. The energy of a given rotameric state for a given residue E_r, depends on its relative population in the database. The total rotameric energy is

$$E_{\mathrm{rot}} = \sum_{i=2}^{N-1} E_r(A_i, \theta_i) \qquad (1e)$$

where the scaling factor for the rotamer energy is 0.5.

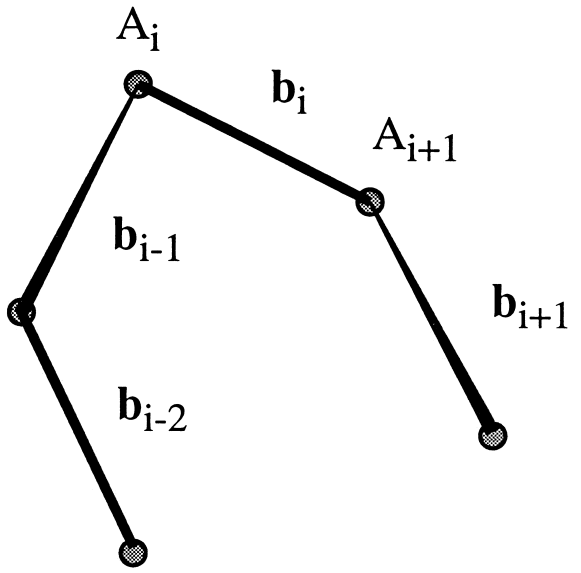

Figure 6. Schematic drawing of the chain showing the various symbols used to define equation (1c). A_i (A_j) is the identity of ith (jth) residue. The $\mathbf{b}_k$ ($k = i-2, i+1$, etc.) are the corresponding bond vectors.

Local side-chain of orientational preferences

We also employ an energetic term that specifies the different relative orientational preferences of neighboring C^{α}-C^{β} vectors. Since the position of the C^{β} atoms are determined from the position of two backbone vectors (Rey & Skolnick, 1992), this contribution to the potential depends only on the backbone conformation; it is, however, amino acid pair-specific. Its functional form is as follows:

$$E_{\beta} = \sum_{i=2}^{N-1} \left(\sum_{k=1}^{4} E_{\beta}^{k}(\cos \alpha_{i,i+k}, A_i, A_{i+k}) \right) \qquad (1f)$$

where $\cos \alpha_{i,i+k}$ is the cosine of the angle between the C^{α}-C^{β} vectors of residues i and $i+k$, A_i, A_{i+k} are the identities of the corresponding residues (see Figure 7). E_{β} is the residue-specific side-chain orientational coupling energy. The scaling factor for this part of the potential is 1.0.

Long-range interactions: burial energy

If the number of side-chain contacts (defined below) for a given residue is greater than a residue-dependent burial

Table 2. Characterization of Properties Associated with the R14 Distribution
A. *Description of the bins for the* R_{14} *distribution*

Bin number	R_{14} value	Chirality	Description
1	< 79	Left-handed	Beta
2	79–55	Left-handed	Beta
3	36–56	Left-handed	Loop
4	26–35	Left-handed	Loop / turn
5	13–25	Left-handed	Helix
6	1–13	Right / Left-handed	Prohibited
7	14–25	Right-handed	Helix
8	26–45	Right-handed	Loop / turn
9	46–64	Right-handed	Loop
10	65–71	Right-handed	Beta
11	72–81	Right-handed	Beta
12	> 81	Right-handed	Non-existent

R_{14} value shows the square of the end-to-end distance for three consecutive vectors in lattice units. The chirality is considered to be left-handed, if for the consecutive vectors $\mathbf{b}_{i-1}$, $\mathbf{b}_i$, $\mathbf{b}_{i+1}$ $(\mathbf{b}_{i-1}\otimes\mathbf{b}_i)\cdot\mathbf{b}_{i+1}$ is less than zero; otherwise it is right-handed.

B. *Energies* $E_{14|14}$ $(R_{14}^{i-1}|R_{14}^{i})$ *for 12 conditional bins of the* R_{14} distribution

Bin number	1	2	3	4	5	6	7	8	9	10	11	12
1	0.0	–2.2	–1.4	0.0	2.0	2.0	0.0	2.0	2.0	2.0	2.0	2.0
2	6.8	–1.8	–0.9	2.4	2.4	2.0	1.6	0.7	0.2	0.1	0.5	6.8
3	5.3	–0.4	–0.8	1.6	1.5	4.2	–1.4	–0.5	–0.7	1.5	3.0	5.0
4	2.0	–0.3	–0.5	2.5	–0.7	2.0	–0.7	–0.5	–1.1	2.3	1.4	2.0
5	2.0	1.6	0.3	2.5	–0.8	2.0	0.4	–1.2	–1.6	3.9	2.0	2.0
6	2.0	0.9	–1.0	1.6	1.6	1.6	–1.9	–0.5	–0.2	2.0	2.0	2.0
7	2.0	2.4	0.4	2.6	3.6	3.4	–2.2	0.0	0.3	5.0	6.2	2.0
8	5.8	–1.2	–1.0	2.1	1.8	4.0	–0.8	–0.2	–0.5	1.1	1.1	2.0
9	4.8	–1.3	–1.2	1.5	1.6	5.9	–0.3	–0.3	–0.1	1.0	1.4	5.9
10	3.3	–1.9	–0.9	1.6	1.5	2.0	2.7	0.9	0.6	0.9	0.3	2.0
11	3.0	–2.0	–0.7	2.6	2.6	2.0	3.7	1.3	0.9	1.4	–0.2	3.7
12	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

threshold, this residue is considered to be buried; for a given residue type, such a conformation is considered to be the zero of burial energy; whereas, when a residue is unburied, it has energy, E_{unbur}. The energies of the unburied states of the 20 naturally occurring amino acids, together with their burial thresholds, are presented in Table 3. The total one-body burial energy is given by:

$$E_{\text{one}} = \sum_{i=1}^{N} E_{\text{unbur}}(A_i, n_{\text{con}}(A_i)) \qquad (1g)$$

where $n_{\text{con}}(A_i)$ describes the number of side-chain contacts for residue i. The scaling factor for the one body term is 0.5.

Table 3. Contact based one body potential and the threshold number of contacts for the burial status and the maximum number of side-chain rotamers

Amino acid	One body energy in the unburied state	Burial contact threshold	Maximum number of side chain rotamers
Gly	–0.4	1	1
Ala	1.8	2	1
Ser	–0.8	3	2
Cys	2.5	3	3
Val	4.2	3	2
Thr	–0.7	4	1
Ile	4.5	4	3
Pro	–1.6	3	1
Met	1.9	5	7
Asp	–3.5	4	5
Asn	–3.5	4	4
Leu	3.8	4	5
Lys	–3.9	5	14
Glu	–3.5	4	11
Gln	–3.5	5	10
Arg	–4.5	5	21
His	–3.2	4	4
Phe	2.8	4	5
Tyr	–1.3	5	4
Trp	–0.9	5	6

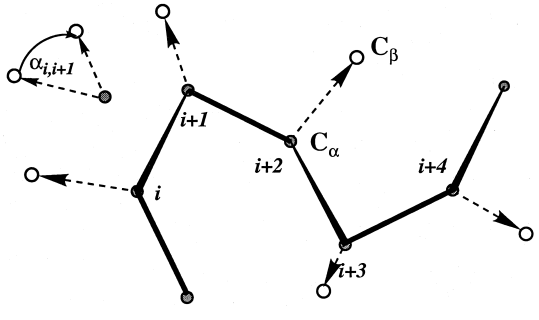

Figure 7. Schematic drawing of the variables appearing in equation (1d) (local side-chain orientational preferences). The open circles represent the positions of the C^β atoms (which are off-lattice), and the shaded circles represent the α-carbon positions. The broken arrows represent C^α-C^β vectors.

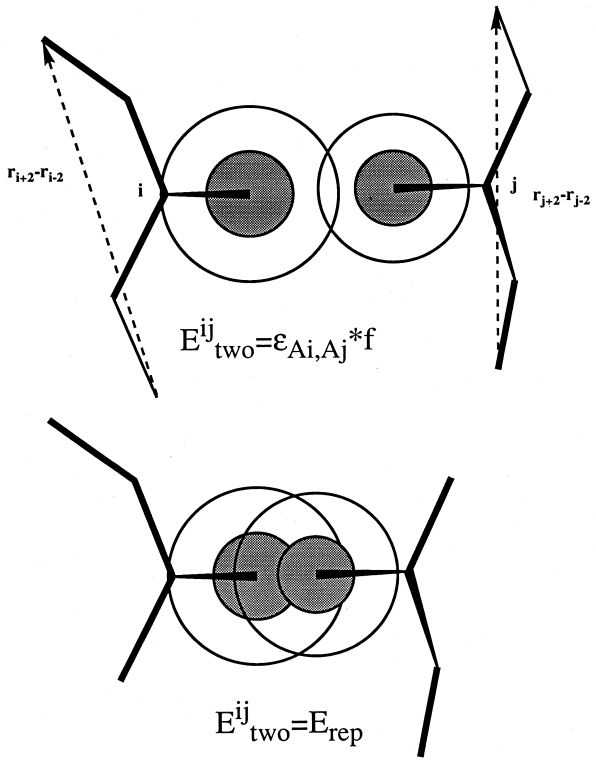

Figure 8. Schematic representation of pairwise interactions. The shaded circles represent repulsive regions of the interaction between the side-chains of residues *i* and *j*. The top panel shows the case when the pair has an attractive basin, and the bottom panel depicts the case when residues *i* and *j* have a repulsive basin. The outer shell represents the basin which may be attractive or repulsive, depending on the identity of the pair of residues. $\mathbf{u}_j$ is the vector connecting residues $j-2$ and $j+2$ and $\mathbf{u}_j = \mathbf{r}_{j+2} - \mathbf{r}_{j-2}$.

Pairwise potential

When the distance between side-chain centers of masses between two residues *i* and *j*, r_{ij}, ($|i-j| > 3$) is less than a residue-dependent threshold value, R_{ij}, residues *i* and *j* are defined to be in contact. The total pair energy of all such contacting pairs is

$$E_{\text{pair}} = \sum_{i=1}^{N} \sum_{j=i+4}^{N} E^{ij}_{\text{two}}(A_i, A_j) \tag{1h}$$

$$E^{ij}_{\text{two}}(A_i, A_j) = \begin{cases} E_{\text{rep}} & \text{if } |r_{ij}| \leqslant R^{\text{rep}}_{ij} \\ a_{ij}\epsilon_{A_i,A_j} & \text{if } R^{\text{rep}}_{ij} \leqslant |r_{ij}| \leqslant R_{ij} \quad \text{and } \epsilon_{A_i,A_j} \geqslant 0 \\ a_{ij} f \epsilon_{A_i,A_j} & \text{if } R^{\text{rep}}_{ij} \leqslant |r_{ij}| \leqslant R_{ij} \quad \text{and } \epsilon_{A_i,A_j} < 0 \end{cases}$$

where the geometry is depicted in Figure 8. R^{rep}_{ij} is the threshold value for the onset of the soft core repulsion. It is amino acid pair-specific and is of the order of $0.4R_{ij}$. E_{rep} is the inner core repulsion value ($6\,kT$), ϵ_{A_i,A_j} is the pair potential between amino acid type A_i and A_j; the 20×20 matrix for ϵ_{A_i,A_j} is shown in Table 4. a_{ij} is a scaling factor for the relative interaction strength of residues *i* and *j*, with $a = 0.6$ if $|i-j| = 5, 6$, and $a = 1$ otherwise. This serves to avoid the problem of non-physical local clustering of side-chains. *f* is a scaling prefactor which depends on the orientation of two backbone fragments containing two interacting residues and is peaked at relative orientations near 20° (the minimum value of *f* occurs at a perpendicular orientation of the interacting chains and is 0.22 of the maximum value) (Kolinski & Skolnick, 1994b). Specifically,

$$f = 1.0 - (\cos^2(\omega_{ij}) - \cos^2(20°)) \tag{1i}$$

$$\cos^2(\omega_{ij}) = \frac{(\mathbf{r}_{i+2} - \mathbf{r}_{i-2}) \cdot (\mathbf{r}_{j+2} - \mathbf{r}_{j-2})}{|\mathbf{r}_{i+2} - \mathbf{r}_{i-2}|\,|\mathbf{r}_{j+2} - \mathbf{r}_{j-2}|}$$

where $\mathbf{r}_k$ ($k = i+2, i-2, j+2, j-2$) is the position of the *k*th α-carbon. This term serves to mimic the side-chain nestling seen in real proteins (preferential packing angle for structural fragments of real proteins). The scaling factor for the pairwise interactions in all coiled coil simulations is 5.0.

Cooperative pairwise interactions

In order to introduce the possibility of a cooperative transition from a molten globule-like state (having poorly defined side-chain packing) to the native state (having well defined side-chain packing), cooperative pairwise interaction templates were introduced. This class of terms attempts to account for the possibility of specific side-chain packing patterns, as well as for interactions which are not readily described in a reduced model. While these terms can magnify the interactions between well defined secondary structure elements, they do not favor any particular kind of secondary structure. The packing template contribution to the potential is given by:

$$E_{\text{tem}} = (\epsilon_{A_i,A_j} + \epsilon_{A_{i+k},A_{j+n}})\delta_{ij}\,\delta_{i+k,j+n}, \quad n = \pm 3, \pm 4 \tag{1j}$$

where δ_{ij}, and $\delta_{i+k,j+n}$ equal 1(0) when two specific pairs of residues *i*, *j* and $i+k$, $j+n$ are (not) simultaneously in contact. The scaling factor for this ''template'' interaction is 4.25.

Total energy of the system

For all of the simulations in the present work, the reduced temperature, T_{red}, (used to determine acceptance ratio of the moves *via* a standard asymmetric Metropolis scheme (Metropolis *et al*., 1953)) was set to 1.85. This temperature was chosen because in the original folding simulations of GCN4, this corresponded to native conditions.

The total energy of the entire system is given by:

$$\begin{aligned} E_{\text{TOT}} &= E_{\text{short}} + E_{\text{long}} \\ &= E_{\text{HB}} + E_{\beta} + 0.25E_{14} + 0.5E_{\text{rot}} + 0.5E_{\text{one}} \\ &= 5E_{\text{pair}} + 4.25\,E_{\text{tem}} \end{aligned} \tag{2}$$

Because the scaling factors of the energy terms used in this study are obtained by parameterization of the model on a generic coiled coil sequence (Hodges' peptides (Hodges *et al*., 1981)) and also worked reasonably well for the GCN4 leucine zipper dimer, we conjecture that the entire energy function might provide a plausible estimation of the relative stability of coiled coils. Whether or not this conjecture is true forms the basis of the investigations described below. We do, however, note that statistical potentials have provided a rationalization of the relative stability of a number of mutations in phage T4 lysozyme (Godzik *et al*., 1993b). Furthermore, the frequency of occurrences of charged residues obeys Coulombs law, albeit with a too high dielectric constant (Bryant & Lawrence, 1991). Thus, the ability of such statistical potentials to provide qualitative insight into relative protein stability is not without precedent.

Table 4. Amino acid pair potential in *kT*

	Bck	Gly	Ala	Ser	Cys	Val	Thr	Ile	Pro	Met	Asp	Asn	Leu	Lys	Glu	Gln	Arg	His	Phe	Tyr	Trp
Bck	–0.4	0.1	0.0	0.1	–0.2	0.3	0.2	0.4	0.5	0.3	0.4	0.3	0.4	0.8	0.8	0.4	0.5	0.4	0.4	0.3	0.3
Gly	0.1	0.5	0.1	0.1	0.3	0.3	0.1	0.4	0.5	0.3	0.1	0.2	0.5	1.0	0.6	0.4	0.5	0.4	0.1	–0.1	–0.3
Ala	0.0	0.1	–0.4	0.1	–0.3	–0.5	0.1	–0.4	0.3	–0.4	0.2	0.2	–0.3	1.0	0.7	0.2	0.6	0.0	–0.5	–0.3	–0.4
Ser	0.1	0.1	0.1	–0.3	0.1	0.3	–0.1	0.4	0.3	0.4	–0.5	–0.1	0.3	0.5	–0.1	0.3	0.3	–0.1	0.1	0.1	0.2
Cys	–0.2	0.3	–0.3	0.1	–3.8	–0.3	0.0	–0.2	–0.1	–0.5	0.5	0.3	–0.3	1.1	0.8	0.3	0.6	–0.2	–0.7	–0.2	–0.1
Val	0.3	0.3	–0.5	0.3	–0.3	–0.6	0.1	–0.5	0.4	–0.3	0.9	0.5	–0.5	1.0	0.7	0.3	0.5	0.3	–0.6	–0.2	–0.5
Thr	0.2	0.1	0.1	–0.1	0.0	0.1	–0.1	0.1	0.3	0.2	–0.2	–0.1	0.3	0.7	0.1	0.1	0.3	–0.2	0.1	0.0	0.2
Ile	0.4	0.4	–0.4	0.4	–0.2	–0.5	0.1	–0.6	0.5	–0.5	0.7	0.7	–0.4	0.9	0.6	0.4	0.4	0.3	–0.6	–0.3	–0.5
Pro	0.5	0.5	0.3	0.3	–0.1	0.4	0.3	0.5	0.4	0.1	0.7	0.4	0.5	1.1	0.8	0.2	0.5	–0.1	–0.1	–0.5	–0.6
Met	0.3	0.3	–0.4	0.4	–0.5	–0.3	0.2	–0.5	0.1	–0.9	0.7	0.4	–0.4	0.9	0.4	0.3	0.4	–0.2	–0.9	–0.5	–0.7
Asp	0.4	0.1	0.2	–0.5	0.5	0.9	–0.2	0.7	0.7	0.7	0.4	–0.2	0.9	–0.4	0.6	0.1	–0.7	–0.4	0.6	0.0	0.3
Asn	0.3	0.2	0.2	–0.1	0.3	0.5	–0.1	0.7	0.4	0.4	–0.2	–0.4	0.6	0.4	0.1	–0.1	0.2	0.0	0.2	–0.1	–0.1
Leu	0.4	0.5	–0.3	0.3	–0.3	–0.5	0.3	–0.4	0.5	–0.4	0.9	0.6	–0.5	1.1	0.7	0.3	0.5	0.0	–0.6	–0.2	–0.5
Lys	0.8	1.0	1.0	0.5	1.1	1.0	0.7	0.9	1.1	0.9	–0.4	0.4	1.1	1.6	–0.4	0.5	1.5	0.7	0.6	–0.1	0.0
Glu	0.8	0.6	0.7	–0.1	0.8	0.7	0.1	0.6	0.8	0.4	0.6	0.1	0.7	–0.4	1.0	0.5	–0.7	–0.1	0.7	0.1	0.2
Gln	0.4	0.4	0.2	0.3	0.3	0.3	0.1	0.4	0.2	0.3	0.1	–0.1	0.3	0.5	0.5	0.4	0.0	0.2	0.0	–0.2	0.0
Arg	0.5	0.5	0.6	0.3	0.6	0.5	0.3	0.4	0.5	0.4	–0.7	0.2	0.5	1.5	–0.7	0.0	0.0	0.0	0.1	–0.3	–0.6
His	0.4	0.4	0.0	–0.1	–0.2	0.3	–0.2	0.3	–0.1	–0.2	–0.4	0.0	0.0	0.7	–0.1	0.2	0.0	–0.2	–0.4	–0.5	–0.6
Phe	0.4	0.1	–0.5	0.1	–0.7	–0.6	0.1	–0.6	–0.1	–0.9	0.6	0.2	–0.6	0.6	0.7	0.0	0.1	–0.4	–1.0	–0.4	–0.6
Tyr	0.3	–0.1	–0.3	0.1	–0.2	–0.2	0.0	–0.3	–0.5	–0.5	0.0	–0.1	–0.2	–0.1	0.1	–0.2	–0.3	–0.5	–0.4	–0.4	–0.3
Trp	0.3	–0.3	–0.4	0.2	–0.1	–0.5	0.2	–0.5	–0.6	–0.7	0.3	–0.1	–0.5	0.0	0.2	0.0	–0.6	–0.6	–0.6	–0.3	–0.5

Protocol for extracting the equilibrium constant from a simulation

In order to compare with experiment, we have to calculate the equilibrium constants associated with the dimer, D, trimer, T, and tetramer, R, species.

$$3\mathrm{D} \leftrightarrow 2\mathrm{T} \tag{3a}$$

$$2\mathrm{D} \leftrightarrow \mathrm{R} \tag{3b}$$

The equilibrium constants (McQuarrie, 1976) are:

$$K_{\mathrm{DT}} = \frac{\{\mathrm{T}\}^2}{\{\mathrm{D}\}^3} \tag{4a}$$

$$K_{\mathrm{DR}} = \frac{\{\mathrm{R}\}}{\{\mathrm{D}\}^2} \tag{4b}$$

with $\{\mathrm{D}\}$, $\{\mathrm{T}\}$ and $\{\mathrm{R}\}$ the concentration of dimer, trimer and tetramer, respectively. The total concentration of individual chains, C_0 (expressed in molecules / Å^3) is given by:

$$C_0 = 2\{\mathrm{D}\} + 3\{\mathrm{T}\} + 4\{\mathrm{R}\}. \tag{4c}$$

Let us define the fraction of the dimers, trimers and tetramers by:

$$x_{\mathrm{D}} = \frac{2\{\mathrm{D}\}}{C_0}; \quad x_{\mathrm{T}} = \frac{3\{\mathrm{T}\}}{C_0}; \quad x_{\mathrm{R}} = \frac{4\{\mathrm{R}\}}{C_0} \tag{5a}$$

Substituting x_{D}, x_{T}, x_{R} to equations (4a), (4b), (4c) we get:

$$K_{\mathrm{DT}} = \frac{8}{9C_0}\frac{x_{\mathrm{T}}^2}{x_{\mathrm{D}}^3} \tag{5b}$$

$$K_{\mathrm{DR}} = \frac{1}{C_0}\frac{x_{\mathrm{R}}}{x_{\mathrm{D}}^2} \tag{5c}$$

$$X_{\mathrm{D}} + X_{\mathrm{T}} + X_{\mathrm{R}} = 1 \tag{5d}$$

Since equations (5b) to (5d) comprise three equations in three unknowns (x_{D}, x_{T}, x_{R}), these equations can be solved numerically (Press *et al*., 1993), once the corresponding equilibrium constants are known. From the statistical mechanical point of view, the equilibrium constants from equations (4a) and (4b), (5b) and (5c) are defined as (Fowler & Guggenheim, 1960); Herschbach, 1959; Mayer & Mayer, 1963):

$$K_{\mathrm{DT}} = \frac{\left(\frac{Z_{\mathrm{T}}}{V}\right)^2}{\left(\frac{Z_{\mathrm{D}}}{V}\right)^3} = \frac{V\left(\frac{VZ_{\mathrm{p,T}}Z_{\mathrm{int,T}}}{\sigma_{\mathrm{T}}}\right)^2}{\left(\frac{VZ_{\mathrm{p,D}}Z_{\mathrm{int,D}}}{\sigma_{\mathrm{D}}}\right)^3} = \frac{Z_{\mathrm{int,T}}^2\sigma_{\mathrm{D}}^3}{Z_{\mathrm{int,D}}^3\sigma_{\mathrm{T}}^2} \tag{6a}$$

$$K_{\mathrm{DR}} = \frac{\{\mathrm{R}\}}{\{\mathrm{D}\}^2} = \frac{Z_{\mathrm{int,R}}\sigma_{\mathrm{D}}^2}{Z_{\mathrm{int,D}}^2\sigma_{\mathrm{R}}} \tag{6b}$$

with V the total volume of the system and σ_γ is the symmetry number (σ = 2!, 3!, 4! for homo dimers, trimers and tetramers, respectively). Z_γ (= $VZ_{\mathrm{p},\gamma}Z_{\mathrm{int},\gamma}$) is the partition function (or phase integral) for species γ (= D, T, R) and is defined as (Herschbach, 1959; Mayer & Mayer, 1963):

$$Z_\gamma = \frac{1}{\sigma_\gamma h^{3N_\gamma}} \times \int \cdots \int \exp(-H/kT)\, dp_1 \ldots dp_{3N}\, dx_1 \ldots dx_{3N} \tag{6c}$$

with H being the total Hamiltonian of the system (the sum of the kinetic and potential energy), h is Planck's constant and $\{p_i\}$, $\{x_i\}$ are the momenta and coordinate degrees of freedom. $Z_{\mathrm{p},\gamma}$ and $Z_{\mathrm{int},\gamma}$ are the integrals corresponding to integration over the momenta and internal coordinate degrees of freedom (also called the internal partition function) for oligomer γ, respectively. Furthermore, there is a factor of V in equations (6a) and (6b) that comes from the coordinate integration over the translational degrees of freedom of the chain (Davidson, 1962; Herschbach, 1959). The integration over the momenta, $Z_{\mathrm{p},\gamma}$, is of the form:

$$Z_{\mathrm{p},\gamma} = \prod_{i=1}^{N_\gamma}\left(\frac{2\pi \mathrm{m}_i kT}{h^2}\right)^{3/2} \tag{6d}$$

Since the number and the mass of atoms remains constant in reactions (3a) and (3b), these contributions cancel.

Let us note a number of facts associated with this approach. The equilibrium constants (equations (4a), (4b), (5b), (5c), (6a) and (6b)) are independent of concentration of the individual chains and are functions of temperature alone. However, the relative populations of the species (x_{D}, x_{T}, x_{R}) are concentration-dependent. Due to the loss of the translational entropy, the lower the concentration C_0 is, the higher is the population of the lower order oligomers. At infinite dilution, only monomers will be present.

Calculation of the internal partition function

The treatment we propose allows one to calculate the internal partition function for any bound state separated from other distinct states by large energy barriers. It is, in principle, exact for any model (continuous or discretized) assuming that adequate conformational sampling is possible. Consider a system comprised of $3N_\gamma$ coordinates. Here, N_γ corresponds to the number of distinct structural elements: in our case, the number of C$^\alpha$ atoms and side-chain centers of mass. The probability of having a conformation (with the first group fixed in space) inside a $3N_\gamma - 3$ dimensional volume element centered about $\mathbf{r} = (\mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}_{N_\gamma})$ is

$$P_u(\mathbf{r}) = \frac{\exp(-E(\mathbf{r})/kT)}{Z_{\mathrm{int},\gamma}}\left(\prod_{i=2}^{N_\gamma} dv_i\right) \tag{7a}$$

$E(\mathbf{r})$ denotes the energy of the internal conformational state $\mathbf{r}$, k is Boltzmann's constant, and T is the temperature. Equation (7a) can be used to precisely calculate the internal partition function $Z_{\mathrm{int},\gamma}$, provided that the corresponding probabilities can be obtained (e.g. from a Monte Carlo simulation).

Now, let us concentrate on the derivation of the $P_u(\mathbf{r})$. We begin the treatment by fixing the origin at the coordinates of the first C$^\alpha$ (Davidson, 1962; Herschbach, 1959). The coordinates of the second C$^\alpha$ are expressed in a spherical coordinate system, (R_2, θ_2, ϕ_2), whose origin is at the first C$^\alpha$. Similarly, the third C$^\alpha$ is expressed in terms of coordinates (R_3, θ_3, ϕ_3) expressed with respect to an origin located at the second C$^\alpha$. The configurational partition function is independent of (θ_2, ϕ_2, ϕ_3), which comprise three Euler angles. The probability of seeing a specific value of (θ_2, ϕ_2, ϕ_3), $P(\theta_2)P(\phi_2)P(\phi_3)$, is just $1/8\pi^2$ (Herschbach, 1959. Thus, $P_u(\mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4, \ldots, \mathbf{r}N_\gamma)$ is given by:

$$P_u(\mathbf{r}_2, \mathbf{r}_3 \ldots \mathbf{r}_{N_\gamma}) = P(\theta_2)P(\varphi_2)P(\varphi_3)P(R_2, R_3, \theta_3 \ldots) = \frac{1}{8\pi^2}P(R_2, R_3, \theta_3 \ldots) \tag{7b}$$

The probability calculated from the Monte Carlo simulation, $P(\mathbf{r})$, has to be corrected for the fact that only a portion, Ω, of the entire range of Euler angles is sampled during the course of the simulation. Thus, $P(\mathbf{r})$ is given by:

$$P(\mathbf{r}) = \frac{1}{\Omega} P(R_2, R_3, \theta_3 \ldots) \tag{7c}$$

Substituting equation (7c) into (7b), gives the probability that the system is free to assume all possible orientations of (θ_2, ϕ_2, ϕ_3):

$$P_u(\mathbf{r}_2, \mathbf{r}_3 \ldots \mathbf{r}_{N_\gamma}) = \frac{\Omega}{8\pi^2} P(\mathbf{r}) = Q_{\mathrm{MC}} P(\mathbf{r}) \tag{7d}$$

For all of the simulations, we calculate Q_{MC} (the correction term for sampling a limited range of rotations) as the average number of observed two consecutive C^α-C^α vectors divided by the total number of possible two consecutive C^α–C^α vectors. In all cases, this number is close to 1/80, and for simplicity, we can assume that our simulations sample a unique orientation of the molecule in space (that is, $Q_{\mathrm{MC}} \approx 1/8\pi^2$).

In the Monte Carlo (Binder, 1986; Metropolis *et al*., 1953) method, the canonical distribution of stages is obtained by Markovian sequence in which the ratio of the probabilities between two conformational states **r** and **r'** are given by:

$$\frac{P(\mathbf{r}')}{P(\mathbf{r})} = \frac{\exp(-E(\mathbf{r}')/kT)}{\exp(-E(\mathbf{r})/kT)} \tag{7e}$$

Thus, by calculating the fraction of time a system spends in a given state (**r**) a dynamic Monte Carlo method provides $P(\mathbf{r})$. From equations (7a) and (7d) we get:

$$Z_{int,\gamma} = 8\pi^2 \exp(-E(\mathbf{r})/kT) \frac{\prod_{i=2}^{N_\gamma} dv_i}{P(\mathbf{r})} \quad \text{with} \quad \mathbf{r} = (\mathbf{r}_2, \ldots \mathbf{r}_{N_\gamma}) \tag{8a}$$

Note that **r** can be any conformational state. However, in what follows, due to the better statistics, the most probable state is used (the state visited most frequently). This approach has the advantage of being exact even for functions having multiple minima on an anharmonic energy landscape (Vieth *et al*., 1995), provided the sampling is efficient. In practice, for systems having substantial conformational fluctuations, the probability $P(\mathbf{r})$ cannot be reliably calculated due to the poor sampling statistics. This requires that a number of simplifying approximations be made.

Local volume factorization

To enrich the sampling, the probability $P(\mathbf{r})$ of the entire structure being in the $3N_\gamma - 3$ dimensional volume element (centered about the most probable conformational state) is approximated as the product of the $N_\gamma - 1$ independent probabilities that each group is in a three-dimensional box centered around its most probable state. That is,

$$P(\mathbf{r}) \cong \prod_{i=2}^{N_\gamma} P_{i,\max(\mathbf{r}_i)} \tag{8b}$$

Equation (8b) is referred to as the local volume factorization approximation. The name comes from the choice of internal Cartesian coordinates that are used to calculate the individual probabilities. The individual probabilities are defined in a similar manner to the total probability:

$$P_{i,\max}(\mathbf{r}_i) = P(\mathbf{r}_{i,\max} - \tfrac{1}{2}\Delta\mathbf{r} < \mathbf{r}_i < \mathbf{r}_{i,\max} + \tfrac{1}{2}\Delta\mathbf{r}); \quad i = 2, N_\gamma \tag{8c}$$

where $\mathbf{r}_{i,\max}$ denotes the most probable position of ith group and $(\Delta r)^3 = dv_i$. The choice of the first bead as the origin of our internal coordinate system is arbitrary, and to remove this arbitrariness, the total probability $P(\mathbf{r})$ (equation (8b)) is better approximated as the product of N_γ independent probabilities divided by their geometric mean:

$$P(\mathbf{r}) \cong \frac{\prod_{i=1}^{N_\gamma} P_{i,\max}(\mathbf{r}_i)}{\left(\prod_{i=1}^{N_\gamma} P_{i,\max}(\mathbf{r}_i)\right)^{1/N_\gamma}} = \left(\prod_{i=1}^{N_\gamma} P_{i,\max}(\mathbf{r}_i)\right)^{1-1/N_\gamma} \tag{8d}$$

(If $P(\mathbf{r})$ were calculated exactly, the results would be independent of the choice of origin). The most probable position of the ith group is computed from the trajectory (in a typical run, we have about 1200 independent structures) as the location ($\mathbf{r}_{i,\max} \pm 1/2\Delta\mathbf{r}$) of maximal frequency of occupation of a given cubic volume element. The idea of the local volume treatment is depicted in Figure 9a and b for the backbone and side-chains, respectively. The volume space is discretized and divided into small boxes, (each side has length Δr = 2.6 Å), with the origin located at the average position of each group. The probability of being in the most frequently occupied box is the fraction of time this box is occupied.

Using the local volume factorization approximation (equation (8d)), the internal partition function defined by equation (8a) for oligomer γ becomes:

$$Z_{\mathrm{int},\gamma} \cong 8\pi^2 \exp(-E(\gamma, \mathbf{r})/kT) \left(\frac{\prod dv_i}{P_{i,\max}(\mathbf{r}_i)}\right)^{1-1/N_\gamma} \tag{9a}$$

$E(\gamma, \mathbf{r})$ corresponds to the energy of the most probable conformation of oligomer γ, and N_γ is the number of groups (united atoms in the lattice protein model) in species γ. While this approximation is not exact, for test energy functions having a similar character as those used here, we have found that the local volume factorization approximation gives satisfactory estimates for the partition functions; errors in the equilibrium constants are on the order of 10 to 20% (Vieth *et al*., 1995). Comparison with exhaustive searches on a number of small systems indicates that the partition function of equation (9a) tends to somewhat overestimate the configurational entropy. Note that equation (9a) is exact for any system with independent groups, regardless of the nature of the energy surface.

For notational convenience let us define for oligomer γ, the average accessible volume per bead (defined here as the geometric mean of the product of all accessible volumes):

$$\langle V_\gamma \rangle = \left(\frac{\prod_{i=1}^{N_\gamma - 1} dv_i}{P(\mathbf{r})}\right)^{1/(N_\gamma - 1)} \cong \left(\frac{\prod_{i=1}^{N_\gamma} dv_i}{P_{i,\max}(\mathbf{r}_i)}\right)^{1/N_\gamma} \tag{9b}$$

In other words, the mean accessible volume is equal to the three-dimensional volume element divided by the average probability of a group being in this volume element.

The internal partition function is defined as the product of the volume elements available to groups in the molecule γ:

$$Z_{\mathrm{int},\gamma} = 8\pi^2 \exp(-E(\gamma,\mathbf{r})/kT) \langle V_\gamma \rangle^{N_\gamma - 1} \tag{9c}$$

Note that $\Pi\, dv_i / P_{i,\max}(\mathbf{r}_i))$ is related to the configurational entropy of all of the groups. The exact treatment (equation 8a)) differs from the approximate treatment (equation (9a)) by definition of the average accessible volume per group. The value of dv_i should be chosen such that the energy $E(\gamma, \mathbf{r})$ remains approximately constant within the

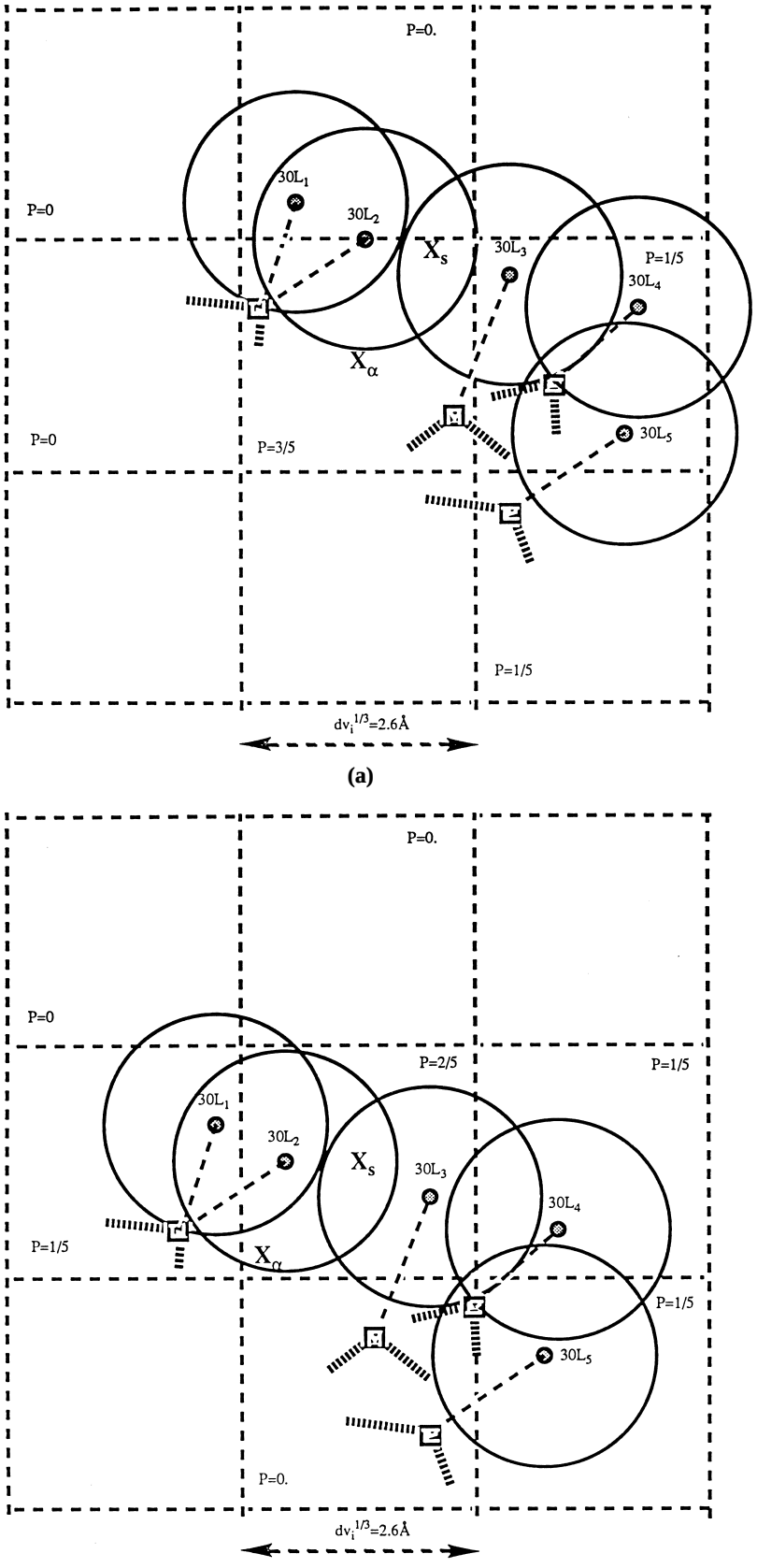

Figure 9. a, The large circles represent the mean radius of interaction for the side-chains of residue Leu30. The small squares depict positions of the centers of mass of the side groups. The small squares depict positions of C^{α} atoms. The broken lines indicate the boundaries of the volume elements (with sides of length 2.6 Å). *P* indicates the probability (fraction of time) of a group (C^{α} or side-chain center of mass) being in the given volume. Each of the five snapshots from the simulation trajectory is indicated by the number of the Leu30 C^{α} (e.g. $30L_2$) or the side-chain center of mass (denoted by X_s for the calculation of the accessible volume for the Leu30 side-chain). Simplified depiction of the calculation of the accessible volume for the C^{α} of Leu30. The probability of the Leu30 C^{α} being in the most probable volume element (which need not be in the same volume element as X_{α}) is 3/5. b, Simplified depiction of the calculation of the accessible volume for the side-chain of Leu30. The average position of the side-chain is placed in the middle of the central box (X_s). In the case presented here, the probability of the side group of Leu30 being in the most probable volume element is 2/5. The origin of the internal coordinate system has been translated with respect to Figure 9a, so that the average position of the Leu30 side-chain center of mass is placed in the middle of the central box (X_s).

corresponding $3N_\gamma - 3$ dimensional volume element. Thus, we chose the box size of $dv_i = (2.6\ \text{Å})^3$ for each accessible volume element. This corresponds to the mean interaction basin for the side-chains and the mean radius for the C^α atoms. The most probable structure $\mathbf{r} = (\mathbf{r}_{1,\max\ldots}, \mathbf{r}_{N_{\gamma,\max}})$ is calculated based on the most frequently occupied position of each of the groups during the simulation. There is no explicit correlation between the group positions that give rise to the most probable state, and this is probably why the specific conformation corresponding to the most probable state was not observed in the simulations. That is why all of the structures having root-mean-square deviation (RMSD) less than $1/2dv_i^{1/3} = 1.3$ Å from the most probable structure $\mathbf{r} = (\mathbf{r}_{1,\max\ldots}, \mathbf{r}_{N\gamma,\max})$ are considered to form the ensemble of structures used to construct the energy $E(\gamma, \mathbf{r})$. The energy of the most probable structure $E(\gamma, \mathbf{r})$ corresponds to the average energy over this ensemble. Nevertheless, the average energy of the structures close in space to the most probable structure is roughly 3.5 *kT* per monomer (see Tables 5 and 8) lower than the average energy of all of the structures. This indicates that the neighborhood of the ''most probable structure' represents a local minimum in the energy landscape and that the free energy calculated using equation (9a) is an approximation to the free energy of a system located in this local minimum. The reason for the choice of the most probable structure as the most probable position of all the independent groups is as follows. For a system having two or more most isoenergetic structures, it is better to choose one of them rather than choose the average over all such structures, as the resulting conformation may in fact reside at a local energy maximum (Vieth *et al*., 1995).

Free energy differences for dimer-trimer-tetramer equilibria

Let us write the difference in free energy for reaction described in equation (3a) (3Dimers → 2Trimers) in terms of the average accessible volume and energy (compare with equation (4a)). First, we define the equilibrium constant for the dimer-trimer equilibrium expressed in the number of molecules (compare to concentration equilibrium constant from equation (4b)) of dimers N_D and trimers N_T:

$$K_{DT}^N = \frac{N_T^2}{N_D^3} = \frac{V^2\left(\frac{N_T}{V}\right)^2}{V^3\left(\frac{N_D}{V}\right)^3} = \frac{\{T\}^2}{V\{D\}^3} = \frac{K_{DT}}{V} \qquad (10a)$$

The free energy difference when dimers form trimers is:

$$\Delta G_{DT}(V, T) = -kT \ln(K_{DT}^N) = -kT \ln(K_{DT}/V) \quad (10b)$$

Substituting equations (6a) and (9c) into equation (10b) we get:

$$\Delta G_{DT}(V, T) = -kT \ln\left(\frac{\sigma_D^3}{8\pi^2\sigma_T^2 V}\frac{\langle V_T\rangle^{2N_T-2}}{\langle V_D\rangle^{3N_D-3}}\right) + \{2E(\mathbf{r}, T) - 3E(\mathbf{r}, D)\} \quad (10c)$$

The factor $k\ln(8\pi^2 V)$ is related to the loss of the translational and rotational entropy when dimers form trimers, whereas the ratio of the accessible volumes describes the change in internal configurational entropy. For a concentration of 2 μM (200 μM), the value of $-kT\ln(8\pi^2 V/2.6\ \text{Å}^3)$ is equal to $-22kT$ ($17.4\,kT$). An equivalent treatment can be done for the dimer/tetramer equilibrium:

$$\Delta G_{DR}(V,T) = -kT \ln\left(\frac{\sigma_D^2}{8\pi^2\sigma_R V}\frac{\langle V_R\rangle^{N_R-1}}{\langle V_D\rangle^{2N_D-2}}\right) + \{E(\mathbf{r}, R) - 2E(\mathbf{r}, D)\} \quad (10d)$$

Generation and equilibration of the starting structures

To generate the starting structures for dimeric, trimeric and tetrameric coiled coils, the initial sequence is aligned to the heptet repeat pattern with the appropriate hydrophobic residues assigned to the *a* and *d* positions. To accomplish this assignment, we use an automated algorithm that calculates the gapless inverse folding score for a given sequence being in each of seven different positions (the alignment can start from *a* through *g* in the coiled coil heptet) in an idealized dimeric coiled coil structure (J. Hirst, M. Vieth, J. Skolnick, C. L. Brooks, unpublished result). The scoring function is the sum of the hydrophobic moment, the sequence-dependent Ramachandran potential, the burial energy and the pair potential described above. The best score per sequence dictates the proper alignment of a given sequence to the heptet repeat. For the tested mutants of the GCN4 leucine zipper, the residue assignment to the *a* to *g* positions based on our automated alignment method is consistent with the Lupas algorithm (Lupas *et al.*, 1991) as well as with the experimental data (Harbury *et al.*, 1993).

Next, the aligned sequence is projected onto a helical wheel. Then, the helices are built on the lattice and translated to be in the neighborhood of each other, with more or less correct registration (Harbury *et al.*, 1993; Vieth *et al.*, 1994a). Equilibration of the trial structure is then initiated by on lattice Monte Carlo simulations. Side-chain contact restraints between the corresponding *a* and *d* residues, along with helical biases for the entire chain backbone, are used to generate appropriate coiled coil configurations. The left-handed supertwist of the helices spontaneously emerges during this part of the equilibration process. The resulting coiled coil conformation is subjected to 180,000 Monte Carlo cycles of unrestrained, isothermal Monte Carlo simulations (using the energy function of equation (2) and a temperature of 1.85 K). The lowest energy structure from the isothermal run is considered to be the starting structure for conformational sampling of a given mutant. We note that significant structural rearrangement can occur in the equilibration phase. For example, the II trimer with incorrectly assigned *a* and *d* positions (purposely shifted by one residue) rearranges to a trimer with correct *a* and *d* positions. The correct initial assignment of interresidue contacts, does expedite the equilibration process.

Protocol for production runs

For each mutant, every oligomer is equilibrated as described above, and then subjected to isothermal production simulations under identical conditions to those used for the equilibrium process ($T = 1.85$ K, energy function of equation (2)). However, to enhance the conformational sampling in the vicinity of the ''native'' conformation, we slightly modify the original Monte Carlo procedure. A given production run consists of 90 iterations,

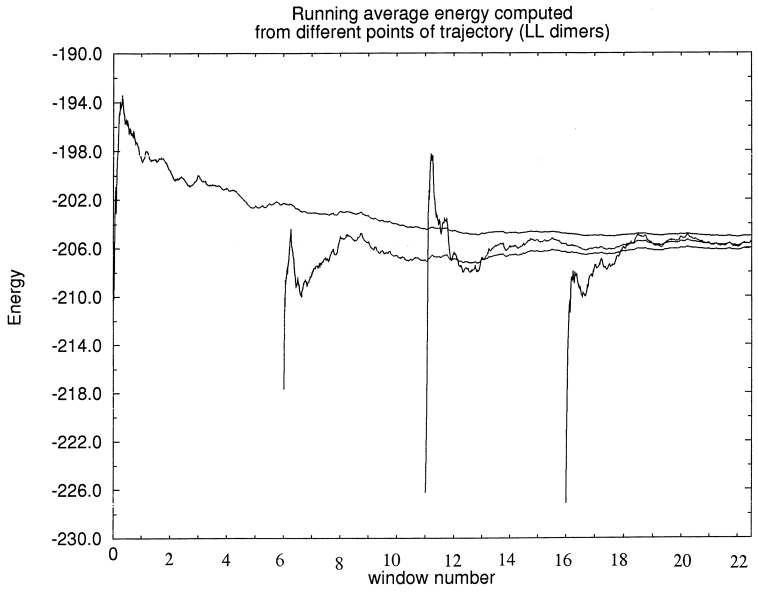

Figure 10. Running average energy showing the parallel dimers of the LL mutant. The four curves show running average energies computed from the different points in the trajectory.

each comprised of 2000 Monte Carlo cycles. For iteration i, the starting structure is chosen to be the lowest energy structure from iteration $i-1$. For each run, a total of 1800 snapshots is collected, i.e. a snapshot is taken every 100 Monte Carlo cycles. Then, a window size of 8000 (80 snapshots) is used to determine the running average energy. The plateau for the running energy is determined by two criteria. First, the running average energies computed at the end of the two consecutive windows (starting from the putative point where the energy reaches a plateau value) are required to be within 1 kT / monomer. Next, the final average energy computed from the point where the plateau is reached and all subsequent windows is also required to be within 1 kT / monomer.

Figure 10 shows four plots of the running average energies computed from four different starting windows. There are a total of 22 possible starting points and the earliest snapshot that fulfills the above criteria is considered to be the starting plateau point. From this point, the average energy as well as the average accessible volumes are recomputed. In most cases, the plateau for the average energy is reached after 80,000 (equivalent to the tenth window) Monte Carlo cycles. In some cases, the plateau is never reached; these runs are discarded. In general, in the plateau region, the average RMSD of a single structure from the average structure is about 1.5 Å. Most structures reside within a 3 Å wide tube centered about the average structure. The average overlap between interchain side-chain contact maps for dimers, trimers or tetramers is 70, 75, 80%, respectively.

For each sequence (wild-type and the seven mutants) in each assumed structure (dimers, trimers, tetramers), a minimum of four production runs were performed. The free energies from four runs are averaged, and if the fifth run does not change the average values of the free energy by more than 1 kT per monomer (or roughly 1%), then the procedure is terminated. Otherwise, the procedure continues until the 1 kT per monomer convergence in the average free energy is reached.

The comparison of the free energies from the ''true'' Monte Carlo procedure with the modified procedure is presented in Table 5 for the LL mutant. The free energies were calculated in the identical manner using the methodology described below. What is clear from Table 5 is that ''true'' Monte Carlo procedure has larger average energy values, larger entropy and larger RMS fluctuations. The differences in free energies between dimers, trimers and tetramers from the ''true'' Monte Carlo results lie within 0.6 kT / per monomer of the values for the modified procedure. Recognizing that we emply a 1 kT / monomer threshold for convergence, it is apparent that the ''true'' Monte Carlo procedure give results similar to the modified Monte Carlo procedure. What is more important is that regardless of the Monte Carlo procedure used, the dominant species computed remain the same over the examined concentration regime. However, only a quarter of the ''true'' Monte Carlo runs exhibit a plateau region in the running average energy and that is why modified Monte Carlo is more ''productive'' for the purposes described in this paper.

Results

The method described in the previous section has been applied to the various oligomeric states depicted in Figure 1 for the mutants of the GCN3 leucine zipper (Harbury *et al.*, 1993). All helical orientations (including three antiparallel orientations of helices in tetramers) were considered for the LL mutant of the GCN4 leucine zipper. As can be seen from Table 6, the computed free energies for the antiparallel species are considerably higher (~ 5 kT / monomer) than those of the parallel species. This energy difference is large enough to allow for the dismissal of antiparallel structures for this mutant. The preferential stability of parallel over antiparallel species seems to be conserved for other tested mutants (Table 6). Because of the invariance of destabilizing charged residues at the e and g positions in the antiparallel arrangement, we would expect this trend to hold for the other mutants as well. Thus, for all other mutants, we only examined the oligomeric equilibria between parallel dimers, trimers and tetramers. However, we ignore the possibility of higher order aggregates (i.e. pentamers, dimers of trimers, etc.), which in principle

Table 5. Comparison of the results from the ''true'' Monte Carlo procedure with those obtained from the Monte Carlo with periodic restarts from the minimum energy structure for the last 6000 cycles for the parallel orientations of II mutant

	Thermodynamic and structural parameters			Configurational free energy difference / k_BT per monomer	
	Dimer	Trimer	Tetramer	$\Delta G_{D\to T}$	$\Delta G_{T\to R}$
LL	–131.7[a]	–136.5	–134.0	–4.8[f]	2.5[g]
MC	–110.9[b]	–118.6	–116.9		
with	–106.4[c]	–116.2	–113.8		
restarts	20.8[d]	17.9	17.1		
	1.68[f]	1.49	1.49		
LL	–141.4	–146.5	–144.6	–5.1	1.9
''true''	–105.9	–116.6	–114.8		
MC	–103.6	–112.1	–113.4		
	35.2	29.9	29.8		
	2.11	1.90	2.19		

[a] The total free configurational energy per monomer expressed in *kT* units: $G_{conf} = (E(\gamma, \mathbf{r}) + (N_\gamma - 1)\ln\langle V_\gamma\rangle - (N_\gamma - M)\ln dv_i)/M$, where $dv_i = 2.6^3$, M is the number of chains and $E(\gamma, \mathbf{r})$ is the energy of the most probable ensemble of structures.
[b] The average energy of ensemble of the most probable structures (per monomer) $E(\gamma, r)$ expressed in *kT* units.
[c] The average energy of all of the structures, computed from Monte Carlo runs.
[d] Configurational entropy contribution of all of the groups (per monomer) calculated using the local volume factorization approximation. In this Table, we report the values of effective entropy, i.e. the effective entropy = $((N_\gamma - 1)\ln\langle V_\gamma\rangle - (N_\gamma - M)\ln dv_i)/M$, where $dv_i = 2.6^3$ and M is the number of chains.
[e] The average RMSD of structures from the average structure for Monte Carlo runs.
[f] The configurational free energy difference per monomer for the reaction D → T.
[g] The configurational free energy difference per monomer for the reaction T → R.

could also occur (Chmielewski, 1994; Harbury *et al.*, 1993).

Employing equations 3 to 10 for each mutant, we calculated the partitioning between dimers, trimers and tetramers. Because of the limited accuracy of our energy function as well as simplifications in the probability calculations, we restrict the calculation to the prediction of the dominant species for each mutant at a given concentration. Thus, we calculate the partioning at low (2 μM) and high (200 μM) concentration (Harbury *et al.*, 1993; see, for example Table 7). For all cases considered, we find that over the experimentally measured concentration regime, the predicted dominant species is the same. However, in all cases, as would be expected because of the law of mass action, in the low concentration regime (about 2 μM), the population of lower order oligomers increases.

Table 6. Configurational free energies per monomer and average RMSD from Monte Carlo runs for different mutants in different studied arrangements

		Parallel orientations			Antiparallel orientations				
a	d	2	3	4	2a	3a	41a	43a	44a
GCN4		121.7[a]	124.3	119.7	118.7	121.2			
		1.69[b]	1.61	1.61	2.05	2.02			
I	L	125.4	129.7	130.1	120.5	123.1			
		1.60	1.60	1.86	1.80	1.85			
I	I	126.5	136.2	133.8					
		1.62	1.54	1.65					
L	I	126.7	133.0	135.6	123.9	126.6		126.9	125.0
		1.67	1.61	1.55	1.79	1.99		2.00	1.61
V	I	122.9	130.4	126.3					
		1.64	1.49	1.44					
L	V	128.3	132.8	129.7					
		1.71	1.57	1.57					
V	L	127.4	133.8	126.1					
		1.63	1.59	1.64					
L	L	131.7	136.5	134.0	126.9	130.3	131.4	132.3	131.8
		1.68	1.49	1.49	1.76	1.88	2.34	2.00	1.80

[a] Configurational free energy per monomer is given by: $G_{conf} = (E(\gamma, \mathbf{r}) + (N\gamma - 1)\ln\langle V\gamma\rangle - (N\gamma - M)\ln dv_i)/M$, where $dv_i = 2.6^3$, M is the number of chains and $E(\gamma, \mathbf{r})$ is the energy of the most probable ensemble of structures.
[b] The average RMSD of structures from the average structure for Monte Carlo runs.

Table 7. Comparison of the simulation prediction of dominant species with experiment

Mutation a	Mutation d	Dominant species from Experiment	Dominant species from Simulation	Concentration dependence from simulation
Wild-type		2	2	99.5:0.5:0[a]
				95:5:0
I	L	2	2, 3	63:37:0
				22:78:0
I	I	3	3	0:100:0
				0:100:0
L	I	4	4	2:46:52
				0:20:80
V	I	?	3	0:100:0
				0:100:0
L	V	3	3	49:51:0
				15:85:0
V	L	(2, 3)	3	2:98:0
				0.5:99.5:0
L	L	3	3	33:67:0
				8:92:0

[a] Top and bottom rows in each cell indicate the percentage of the dimers, trimers and tetramers at 2 μM and 200 μM concentration of a peptide, respectively.

A comparison of the predictions with the experimentally determined degree of chain association is presented in Table 7. In five out of eight cases, the predictions are in complete agreement with the experimentally determined dominant species. In the offending case of the IL mutant, trimers and dimers are assigned to be the dominant species. Experiment indicates that only dimers are present. This may reflect the inaccuracy of the potential as well as accumulations of errors in the entropy calculation (for this mutant, the entropy increases with degree of association). In the case of the VL mutant, dimeric species make a negligible contribution, and trimers are assigned to be the dominant species over the entire concentration regime. In contrast, experiment indicates that both dimers and trimers are populated (Harbury *et al.*, 1993). For the VI mutant, trimers are predicted to be the only species over the entire concentration regime, whereas experiment shows that multiple species are populated (Harbury *et al.*, 1993).

In Table 8, the individual contributions to the free energy per monomer are shown for every mutant investigated. The contributions of the different energy terms to the total average is shown in Table 9. From Table 8, it is apparent that in all cases the dominant contribution to the effective entropy change (60 to 90%) comes from the side-chains; the effective entropy change for the backbone is smaller, but non-negligible. This substantial contribution of

Table 8. Dissection of free energy contributions to stability

		–ENERGY / k_B T per monomer			Effective ENTROPY / k_B per monomer		
a	d	Dimer	Trimer	Tetramer	Dimer	Trimer	Tetramer
GCN4		100.0[a]	103.7	99.0	21.7[d]	20.6	20.7
		96.2[b]	100.6	94.9	11.6[e]	10.9	11.0
		41%[c]	56%	59%			
I	L	106.4	109.5	108.8	19.0	20.2	21.3
		103.9	106.4	105.1	10.1	10.6	11.1
		45%	56%	60%			
I	I	106.0	117.2	115.9	20.5	19.0	17.9
		103.3	114.6	112.3	10.9	10.0	9.3
		47%	63%	64%			
L	I	106.4	113.4	118.1	20.3	19.6	17.5
		101.7	109.5	114.8	10.8	10.3	9.1
		42%	56%	61%			
V	I	102.2	113.4	109.5	20.7	17.0	16.8
		98.4	111.0	107.0	11.0	8.9	8.8
		47%	60%	61%			
L	V	106.4	113.8	111.8	21.9	19.0	17.9
		101.9	110.8	108.3	11.7	10.0	9.3
		45%	58%	61%			
V	L	107.6	114.2	106.6	19.8	19.6	19.5
		103.9	110.8	102.5	10.5	10.3	10.2
		44%	57%	59%			
L	L	110.9	118.6	116.9	20.8	17.9	17.1
		106.4	116.2	113.8	11.0	9.4	8.9
		44%	57%	60%			

[a] The average energy of the ensemble of the most probable structures (per monomer) $E(\gamma, \mathbf{r}) / M$.

[b] The average energy of all of the structures, computed from Monte Carlo runs.

[c] Percentage of the total energy, that is long-range (pairwise, templates, contact one body).

[d] Configurational entropy contribution of all of the groups (per monomer) calculated using the local volume factorization approximation. In this Table, we report the values of effective entropy, i.e. the effective entropy = $((N\gamma - 1)\ln\langle V\gamma\rangle - (N\gamma - M)\ln dv_i) / M$, where $dv_i = 2.6^3$ and M is the number of chains.

[e] The side-chain effective entropy in local volume factorization approximation.

Table 9. Dissection of the average energy per monomer for parallel dimers, trimers and tetramers

a	d	Dimer −E / monomer	Trimer −E / monomer	Tetramer −E / monomer
GCN4		**96.2**[a]	**100.6**	**94.9**
		14.9[b]; 8.7[c]; 18.7[d]	25.3; 13.6; 17.0	24.0; 15.0; 15.9
		39.6[e]; 15.9[f]; –5.1[g]	30.9; 16.9; –6.4	27.1; 16.9; –7.2
I	L	**103.9**	**106.4**	**105.1**
		19.0; 10.6; 19.1	24.9; 17.9; 17.4	27.9; 18.6; 16.3
		41.3; 14.9; –5.6	32.8; 16.6; –7.1	28.7; 16.3; –6.4
I	I	**103.3**	**114.6**	**112.3**
		21.4; 12.4; 18.9	31.4; 23.8; 17.3	33.6; 22.0; 16.4
		37.0; 14.9; –6.0	27.5; 16.7; –6.0	26.6; 16.2; –6.0
L	I	**101.7**	**109.5**	**114.8**
		17.6; 10.7; 18.6	26.5; 18.3; 16.9	31.9; 22.0; 17.2
		41.0; 14.8; –5.1	33.0; 16.4; –5.9	30.9; 16.3; –6.6
V	I	**98.4**	**111.0**	**107.0**
		19.5; 11.4; 18.6	29.4; 20.6; 17.4	29.1; 19.6; 17.0
		35.6; 14.9; –5.3	28.7; 16.7; –5.6	27.3; 16.5; –5.3
L	V	**101.9**	**110.8**	**108.3**
		19.6; 11.5; 18.4	28.3; 19.8; 17.1	31.3; 18.0; 16.3
		38.8; 15.2; –5.3	31.3; 16.7; –6.0	29.1; 16.5; –6.2
V	L	**103.9**	**110.8**	**102.5**
		19.4; 11.4; 18.9	27.9; 18.5; 17.5	25.8; 18.5; 16.4
		40.2; 15.1; –4.5	32.3; 16.6; –5.6	28.5; 16.3; –6.4
L	L	**106.4**	**116.2**	**113.8**
		20.3; 12.1; 18.5	30.2; 18.8; 17.4	30.9; 20.5; 16.7
		42.8; 15.0; –6.2	36.8; 16.4; –7.0	32.8; 16.6; –7.3

[a] The total energy per monomer.
[b] Pairwise.
[c] Template.
[d] Hydrogen bond.
[e] Side-chain orientational correlations.
[f] Contact one body.
[g] R_{14} energy.
Note: side-chain rotamer energy is in all cases close to $-4kT$ / monomer.

the backbone to the overall internal entropy change may be related to the slightly exaggerated mobility of the backbone in our model with respect to the side-chain mobility. The trend that the internal entropy of the backbone decreases with the increased degree of oligomerization is intuitively reasonable, but in real systems, the magnitude of the internal entropy change is probably smaller, and the dominant contribution to the internal entropy change comes from the side-chains (Novotny *et al.*, 1989). Since the exposed surface area is larger for the lower order oligomers, these structures possess more exposed side-chains, whose effective configuration entropy is larger. The greatest contribution to the entropy (largest accessible volume) comes from the C-terminal ends of the molecules. This prediction is consistent with the crystal structures of the wild-type dimer and the LI tetramer. In both cases, the last two C-terminal residues are highly disordered (Harbury *et al.*, 1993; O'Shea *et al.*, 1991).

In the case of the wild-type GCN4 leucine zipper, trimers are favored energetically (we use a constant volume, number particles and temperature ensemble) and disfavored entropically. In the wild-type, trimers are more stable by about 2 kT / monomer than dimers, but over the experimental concentration regime only dimers are predicted. A plot of the energy per residue is presented in Figure 11 for the wild-type dimer (a) and trimer (b). In addition, the N16V mutant (= VL) is also depicted. It is interesting to note that according to the calculation of the energy per residue, Asn in fact destabilizes the trimer locally (residues 14 to 18) (by 6.1 kT per monomer plus a constant value that reflects the effect of the mutation on the unfolded state). The local destabilization of Asn in the dimer is smaller (by 4.1 kT per monomer plus a constant value that reflects the effect of the mutation on the unfolded state) for residues 14 to 18. Other parts of the wild-type trimer play a stabilizing role, and thus, compensation effects are present. These calculations indicate that the effect of a single point mutation is not local, but propagates for at least one helical turn. This is in agreement with studies (Holtzer & Holtzer, 1990) on tropomyosin fragments, where compensation effects are also present. We find that the N16V mutation stabilizes trimers more than dimers by roughly 2.8 kT per monomer. This number can be obtained by subtracting the difference in free energy of trimers of the VL mutant and wild-type from the difference in free energy dimers of VL and wild-type (data from Table 6).

We observe that trimers and tetramers prefer Leu in the *a* position. Ile in this position favors trimers and destabilizes tetramers. For the seven mutants of

the GCN4 leucine zipper, Figure 12a (b) shows a plot of the total energy difference between dimers and trimers (trimers and tetramers) as a function of the sum of the pair energy, template energy and local-side chain orientational energy difference and the pair energy difference alone. The sum of the three above terms is highly correlated (with a correlation coefficient, $R = 0.99$) with the total difference in energy. This suggests that three specific terms (pair energy, template energy and local side-chain orientational energy) fully determine the energetics of the switch between two, three and four-stranded coiled coils. If one were to assign the single most important energetic contribution to the switch based on our simulations, then the pairwise energy would be the most important, since this is the only term which in and of itself is well correlated ($R = 0.92$ to 0.94) with the total energy change (see Figure 12a and b). The correlation coefficient of the template energy change with the total energy change is 0.60 in the dimer-trimer equilibria, and 0.74 for trimer-tetramer equilibria. The corresponding values for side-chain orientational preferences are 0.74 and 0.22, respectively.

Our calculations also suggest that short-range, intrinsic secondary structure preferences (hydrophobic moment) favor lower order oligomers. Furthermore, the reduction in side-chain effective entropy on burial in the core of trimers and tetramers also acts to favor lower order oligomers (Table 8). The long-range interactions (burial preferences, cooperative side-chain packing interactions, and side-chain

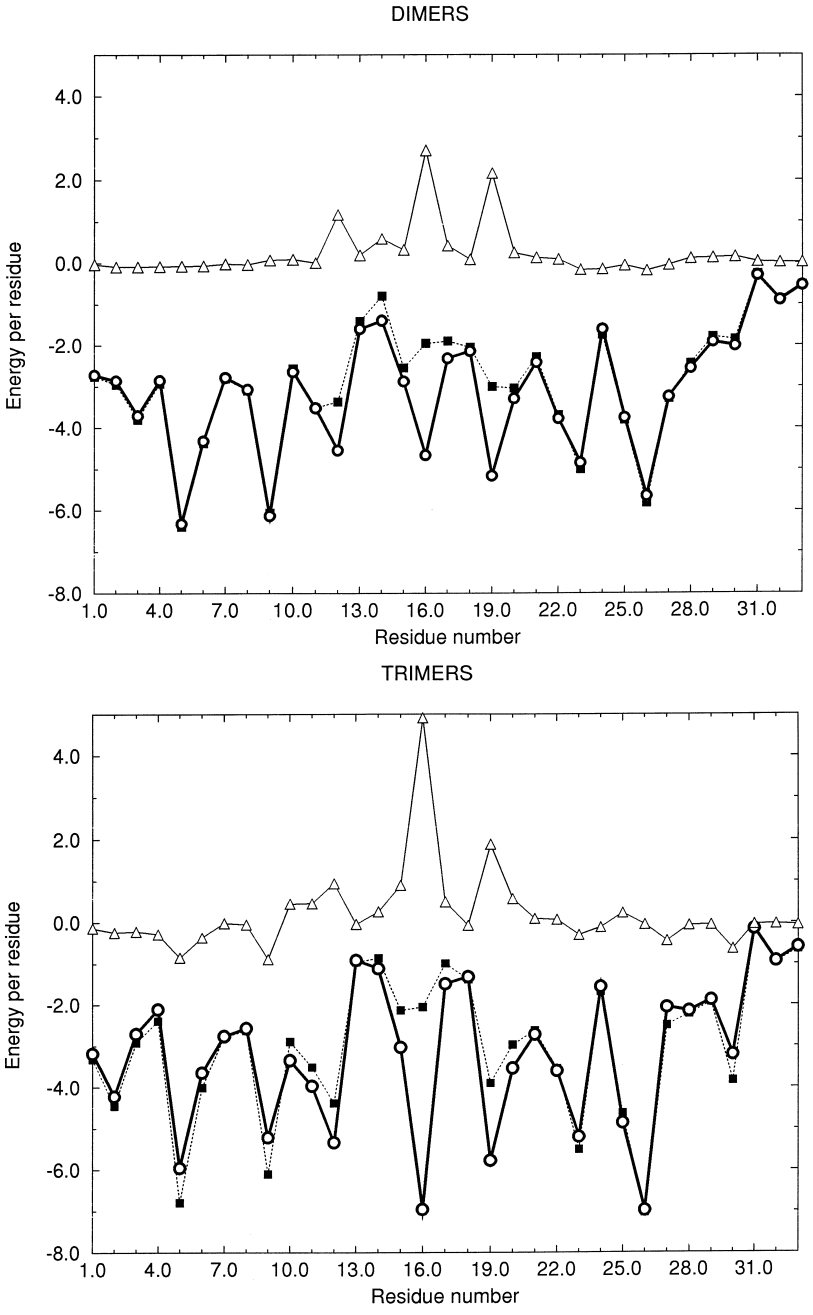

Figure 11. a, The plot of the energy per residue for dimers. The single point mutation V16N destabilizes the dimer from residues 12 through 19. b, The plot of the energy per residue for trimers. The single point mutation V16N destabilizes trimers from residues 10 to 20, but compensation effects are present. Residues 5, 9, 23 and 29 are, in fact, more stable in the wild-type trimer. The open circles represent the VL mutant, the filled squares indicate the wild-type mutant, and the open triangles represent the energy difference per residue between the wild-type mutant and the VL mutant.

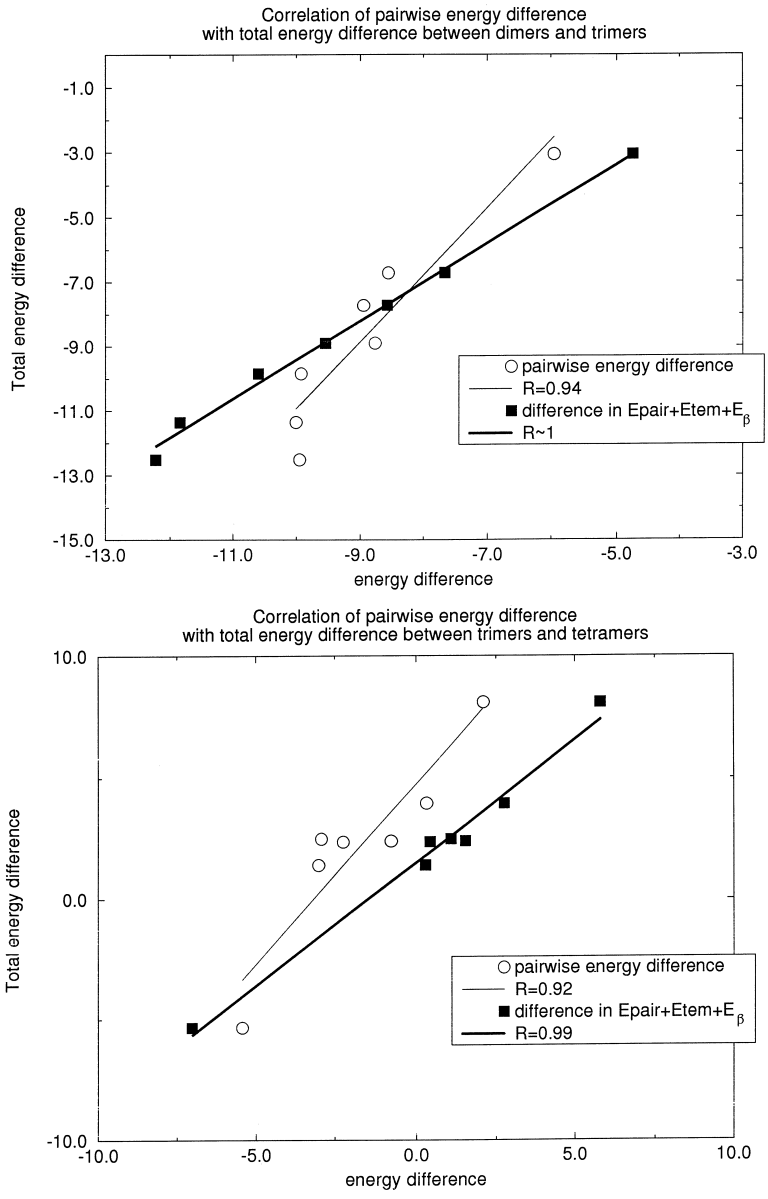

Figure 12. a, Correlation between the energy difference per monomer between dimers and trimers with differences coming from various energy terms. The filled squares present the total energy difference as a function of the sum of the pair term, template term, and hydrophobic moment term. The open circles represent the total energy difference as a function of the pair energy difference alone. b, Correlation between the energy difference per monomer between trimers and tetramers with differences coming from various energy terms. The filled squares present the total energy difference as a function of the sum of the pair term, template term, and hydrophobic moment term. The open circles represent the difference as a function of the pair energy difference alone.

pairwise interactions, the last being the most specific) favor higher order oligomers. In higher order multimers, side-chains in the core (*a* and *d* residues) are more buried and experience additional favorable hydrophobic interactions. The competition between short-range and long-range interactions and the effective side-chain entropy change are the major factors that determine the dominant species for the mutants studied here.

Harbury *et al.* (1993) have explained the different stabilities of various GCN4 mutants based on the preferential relative angular packing of different side-chains. In the known crystal structure, parallel packing occurs at the *a* positions in tetramers and *d* positions in dimers, whereas perpendicular packing occurs at the *a* positions in dimers and *d* positions in tetramers. Acute packing is exhibited by trimers. They argue that Ile and Val side-chains prefer to pack in the perpendicular or acute fashion, and Leu in the parallel fashion. This is perhaps the reason why LI forms trimers, IL tetramers and II trimers. In our model, however, the related term (see rotamer population, Table 10) does not exhibit such a trend. Our explanation of specificity is based on the competitive effects of the pairwise interactions, side-chain packing (long-range interactions favor higher order species) and side-chains orientational packing preferences E_β (short-range interactions favor lower order oligomers) together with the loss of configuration entropy (which favors lower order oligomers). As indicated in Tables 8 and 9, the changes in other energetic and entropic terms upon change of the association state show trends that do not depend strongly on sequence. It is interesting, however, to analyze the Leu and Ile rotamer population in the *a* and *d* positions in various oligomers. Table 10 shows the simulation results for the frequencies of occurrence of the three most populated database rotamers of Leu and Ile in the interfacial positions. For the *a* position, in 11 out of 12 cases the most statistically populated rotamer in the database is also the most populated in our

Table 10. Frequency of the three most populated rotamers observed in the simulations for IL, II, LI and LL mutants

		Dimers		Trimers		Tetramers	
a	d	a	d	a	d	a	d
I	L	**98**	45[a]	**92**	44[a]	**91**	**47**
		2	**53**	7	**49**	5	38
		0	0	0	7	4	14
I	I	**99**	**96**	**94**	**84**[a]	93	61
		1	4	5	12	3	30
		0	0	1	4	4	8
L	I	**60**	**91**	**72**	**95**	48	**56**[a]
		39	9	26	5	**51**	30
		1	0	2	0	0	14
L	L	**62**	**57**	**65**	36[a]	**65**	44
		38	36	30	**56**	35	**48**
		0	2	5	8	0	8

The most populated rotamer is shown in bold and the rotamers are presented in decreasing population in the database.
[a] The predicted equilibrium species.

simulations (the exception is Leu in the *a* position in LI tetramer). For the *d* position, in eight out of 12 cases, the statistically most populated rotamer is also the most populated in our simulations. All five exceptions are caused by the Leu residue. In our simulations, Leu exclusively populates the second most statistically populated rotamer in the *d* position in trimers. Thus, on average we see the population of the statistically most favorable rotamer in the majority of species, and there is no selection based on the lowest energy of rotamers. This observation is inconsistent with suggestion of Harbury *et al.* (1993). However, in reduced protein models, this inconsistency may be due to the fuzziness of the simplified representation of side-chains.

Conclusion

In this paper, we have presented a new application of the Mayer & Mayer (1963) approach to calculate the equilibrium constant between dimeric, trimeric and tetrameric coiled coils. This approach combined with a lattice protein model was used successfully to predict the state of association of different mutants of the GCN4 leucine zipper. Local interactions were found to stabilize lower order oligomers, whereas tertiary/quaternary interactions stabilize higher order oligomers. In most cases, the internal entropy of side-chains was found to stabilize low order oligomers. The main differences in the population of different oligomeric species of various mutants come from the interplay between different interaction environments for the *a* and *d* positions in dimers, trimers and tetramers, differential packing preferences and the effective entropy change associated with side-chain burial.

Acknowledgements

We thank Drs A. Godzik, M. Milik, K. Olszewski, S. Debolt and F. Sheinerman for helpful discussions. We also thank Professors R. Baldwin, Jan Hermans and anonymous referees for suggesting clarifications of some points presented in this paper. This work was supported in part by NIH grants GM-38794, GM-37554 and FIRCA PA-91-77.

References

Banner, D. W., Kokkinidis, M. & Tsernoglou, D. (1987). Structure of the ColE1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657–675.

Binder, K. (1986). *Monte Carlo Methods in Statistical Physics*. Springer Verlag, Berlin.

Bryant, S. H. & Lawrence, C. E. (1991). The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential. A statistical model for non bonded interactions. *Proteins: Struct. Funct. Genet.* **9**, 108–119.

Chmielewski, J. (1994). General strategy for covalently stabilizing helical bundles. A novel five-helix bundle protein. *J. Am. Chem. Soc.* **116**, 6451–6452.

Cohen, C. & Parry, A. H. (1990). α-Helical coiled coils and bundles: How to design an α-helical protein. *Proteins: Struct. Funct. Genet.* **7**, 1–15.

Cohen, C. & Parry, D. A. D. (1986). α-Helical coiled coils—a widespread motif in proteins. *Trends Biochem. Sci.* **11**, 245–248.

Crick, F. H. C. (1953). The packing of α-helices: simple coiled-coils. *Acta Crystallog.* **6**, 689–697.

Davidson, N. (1962). *Statistical Mechanics*. McGraw-Hill Book Company, Inc., New York.

Ferre-D'Amare, A. R., Prendergast, G. C., Ziff, E. B. & Burkey, S. K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature*, **363**, 38–45.

Fowler, R. & Guggenheim, E., A. (1960). *Statistical Thermodynamics*. Cambridge University Press, Cambridge.

Godzik, A., Kolinski, A. & Skolnick, J. (1993a). Lattice representation of globular proteins: how good are they? *J. Comp. Chem.* **14**, 1194–1202.

Godzik, A., Kolinski, A. & Skolnick, J. (1993b). De Novo and inverse folding predictions of protein structure and dynamics. *J. Comp. Aided Mol. Des.* **7**, 397–438.

Harbury, P. B., Zhang, T., Kim, P., S. & Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science*, **262**, 1401–1407.

Herschbach, D. R. (1959). Molecular partition function in terms of local properties. *J. Chem. Phys.* **31**, 1652–1661.

Hodges, R. S., Saund, A. S., Chong, P. C. S., St-Pierre, S. A. & Reid, R. E. (1981). Synthetic model for two-stranded α-helical coiled coils. *J. Biol. Chem.* **256**, 1214–1224.

Holtzer, M. E. & Holtzer, A. (1990). α-Helix to random coil transition of two-chain coiled coils: experiments on the thermal denaturation of isolated segments of αα-tropomyosin. *Biopolymers*, **30**, 985–993.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kolinski, A. & Skolnick, J. (1994a). Monte Carlo simulation of protein folding II. Application to protein A, Rop, and crambin. *Proteins: Struct. Funct. Genet.* **18**, 353–366.

Kolinski, A. & Skolnick, J. (1994b). Monte Carlo simulation of protein folding I. Lattice model and interaction scheme. *Proteins: Struct. Funct. Genet.* **18**, 338–352.

Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **114**, 181–293.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Mayer, J. E. & Mayer, M. G., (1963). *Statistical Mechanics*. John Wiley & Sons, Inc. New York

McLachlan, A. D. & Stewart, M. (1975). Tropomyosin coiled-coil interactions: evidence for an unstaggered structure. *J. Mol. Biol.* **98**, 293–304.

McQuarrie, D. A. (1976). *Statistical Mechanics*. Harper's Chemistry Series, Harper & Row, New York.

Metropolis, N. A., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–92.

O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two stranded, parallel coiled coil. *Science*, **254**, 539–544.

Novotny, J., Bruccoleri, R. E., & Saul, F. A. (1989). On the attribution of binding energy in antigen-antibody complexes McPC 603, D1.3, and HyHEL-5. *Biochemistry*, **28**, 4735–4749.

Phillips, G. N., Jr, Fillers, J. P. & Cohen, C. (1986). Tropomyosin crystal structure and muscle regulation. *J. Mol. Biol.* **192**, 111–131.

Press, H., Teukolsky, S. S., Vetterling, W. T. & Flannery, B. B. (1993). *Numerical Recipes.* Cambridge University Press, Cambridge.

Ray, A. & Skolnick, J. (1992). Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J. Compt. Chem.* **13**, 443–456.

Vieth, M., Kolinski, A., Brooks, C. L. III & Skolnick, J. (1994a). Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237**, 361–367.

Vieth, M., Skolnick, J., Kolinski, A. & Godzik, A. (1994b). Lattice parameter set available *via* anonymous ftp from ftp.scripps,edu in pub/(MCDP.info,MCDP.Z).

Vieth, M., Kolinski, A. & Skolnick, J. (1995). A simple technique to estimate partition functions and equilibrium constants from Monte Carlo simulations *J. Chem. Phys.* **102**, 6189–6193.

Edited by F. E. Cohen

(Received 18 August 1994; accepted in revised form 24 May 1995)